PARAMETRIC ESTIMATION:

CHOOSING BETWEEN NARROW AND WIDE MODELS

Kjetil Kåresen

May 29, 1992

Cand. Scient. thesis in Statistics
Department of Mathematics
University of Oslo

Acknowledgments

I hereby wish to thank Professor Nils Lid Hjort. I am very grateful for the efficient way he has supervised my Cand. Scient. thesis. First by having the original idea to the project, and then by good advice in all the various stages leading to the final product. He has repeatedly surprised me by his extraordinary mathematical intuition and broad knowledge of statistical methodology; usually having the answer ready almost before I could finish the question.

I also want to thank my wife Beate. Primarily for her patience and support during the long days of hard work, but also for reading a version of the manuscript and giving valuable linguistic advice.

Contents

Chapter 1

Introduction

1.1 Introduction and abstract

The origin of this Cand. Scient. thesis is Nils Lid Hjort's 1991 article "Estimation in moderately misspecified models", to be denoted EMMM for short. In his article Hjort proposes and studies a large sample method to compare estimation based on certain "wide" and "narrow" parametric models. The simplest situation studied by Hjort is the following: We have independent and identically distributed data. To describe the common distribution of the data, we have two competing parametric models. A "wide" model, corresponding to a density $f(y|\theta,\gamma)$, and a "narrow" sub-model corresponding to the density $f(y|\theta,\gamma^0)$, where γ and θ are parameter vectors and γ^0 is some known value. The wide model could for example be a gamma density, and the narrow model could be the exponential density obtained by setting the first parameter of the gamma density equal to one.

The true distribution of the data is considered to correspond to some parameter point of the wide model, and maybe, but not necessarily, also to some point of the narrow model. The problem at hand is to estimate some functional μ of the true density. Since the wide model is assumed correct, the estimand can be written as $\mu(\theta,\gamma)$.

Consider now a statistician facing such an estimation problem. A conservative approach would be to use maximum likelihood estimation based on the wide model, thereby avoiding any biasing of the estimator due to false model assumptions. But if he is quite convinced that the narrow model is true, or almost true, it would be tempting to base the maximum likelihood estimation on the narrow model instead. This would hopefully result in an estimator with reduced variance, making up for a possible bias introduced by employing slightly wrong model assumptions. To make a rational choice in this situation, it could be reasonable to compare the mean square error of the two estimators as a function of θ and γ.

An exact study of this problem in any generality seems to be a hopeless undertaking, thereby making it natural to consider large sample approaches.

If the true model is held fixed and the number of data is tending to infinity, it is easy to show that the wide model ML-estimator will be asymptotically better than the narrow one for all $\gamma \neq \gamma^0$. But this result really only stresses the fact that the optimal choice of

model will also depend on the amount of data available. The degree of misspecification the narrow model can tolerate in order to compare favourably with the wide model, diminishes as n grows.

A large sample approach that should produce nontrivial results must somehow take this fact into consideration. The idea of Hjort is to consider a "true" model that shrinks towards the narrow model as n grows. As it turns out, the exact factor that gives nontrivial limit distributions is to let the "true value" of γ be given by $\gamma^0 + \delta/\sqrt{n}$, for some fixed quantity δ.

In this framework it is possible to derive limit distributions for the narrow and wide ML-estimators and to give simple criteria for when the narrow estimator is better than the wide one. This was done by Hjort in the case of a one-dimensional γ.

In Chapter 2 we shall give generalizations of his results to the case of a vector-valued γ and study a number of new features characteristic of the multi-dimensional case.

Chapter 3 about compromise estimators discusses a class of estimators that try to combine the advantages of the narrow and wide ML-estimator. The idea is to form convex combinations of the two estimators, with the data themselves determining the weight given to each model.

Chapter 4 discusses Bayes estimators and compare them with the narrow and wide ML-estimator. The comparison will be made both from a classical and from a Bayesian point of view.

Although we shall mostly follow a different route of presentation and line of proofs, a number of results in Chapters 2 and 3 will be generalizations of corresponding results in EMMM. Hjort has chosen to give his results in an informal style, omitting or only indicating a number of proofs. As a special case our presentation will thus give rigorous proofs of the main results in EMMM. In particular we shall give a sufficient set of regularity conditions to guarantee the necessary results.

In Chapter 4 we shall also allow ourselves a somewhat less rigorous style of presentation, omitting for example the exact regularity conditions for certain remainder terms to go to zero. Chapter 4 mainly deals with topics not considered in EMMM. The reader should have a knowledge of the most basic facts about Bayes estimation to read this chapter.

In addition to the general theory we shall give a number of examples, illustrating characteristic points of the theory. Exact computations will also be performed in many cases, thus providing an idea of the quality of the large sample approximation.

The mathematics program package Mathematica® has been used for numeric and some symbolic computations in connection with the examples.

We shall confine ourselves to live in the world where the wide model is certainly true. The usually more realistic situation where also the wide parametric model is slightly false, is not considered.

In addition to the already mentioned paper by Hjort, EMMM, we recommend reading Hjort (1991b). In this paper Hjort treats the problems outlined above in the special case where the narrow model corresponds to assuming normality (for example in regression models) and the wide model corresponds to t-ness. This particular problem can not be treated simply as a special case of the general theory because the parameter points of the narrow model are not inner points of the parameter space.

Of other papers treating related topics we mention Bickel (1984) and Berger (1982).

1.2 Definition of the model

We will now give a more precise definition of our model. Let the data, $Y_{ni}, i = 1, 2, \ldots, n$ be random variables, possibly vector-valued. (We use the double subscript on the variables to stress that the distribution depends on n. Later on the first subscript will sometimes be omitted when it does not convey useful information.) The two competing models are given by the wide density $f(y, \theta, \gamma)$ and the narrow density $f(y, \theta, \gamma^0)$. The "true" density of $Y_{n1}, \ldots, Y_{nn}$ is $f(y, \theta, \gamma)$, where $\gamma = \gamma^0 + \delta/\sqrt{n}$. The dimension of θ and γ will respectively be denoted by p and q, and the sum of $p + q$ by r.

Note that in order to economize the use of subscripts, θ and γ denotes both a general parameter point and the point corresponding to the true model. (This is commonly practiced by many authors. The context will make the meaning clear.)

Our notation will indicate that f is a density with respect to Lebesgue measure, but all our results will be equally valid for an arbitrary sigma-finite dominating measure.

For notational simplicity we shall also introduce a variable Y distributed according to the "null"-density $f(y|\theta, \gamma^0)$. It will further be convenient to have a notation for the assembly of θ and γ: $\xi = (\theta', \gamma')'$ and $\xi^0 = (\theta', \gamma^{0\prime})'$. The ML-estimators based on the wide, respectively narrow model, will be denoted by $\hat{\xi}_{\text{wide}}$ and $\hat{\theta}_{\text{narr}}$. The narrow ML-estimator of γ is considered to be γ^0. Correspondingly we define the narrow ML-estimator of ξ as $\hat{\xi}_{\text{narr}} = (\hat{\theta}'_{\text{narr}}, \gamma^{0\prime})'$.

In the following we shall repeatedly refer to the "likelihood"-function, which (of course) is the simultaneous density of $Y_{n1}, Y_{n2}, \ldots, Y_{nn}$. It will be denoted by $L(\xi)$. The logarithm of the likelihood-function is termed the log-likelihood and denoted by $l(\xi)$.

1.3 Mathematical conventions

As we shall usually deal with multivariate quantities, most symbols introduced will be vectors or matrices. The dimensions will not be stated when they are evident from the context. one-dimensional matrices and scalars are identified. We shall often mix matrix multiplication and scalar multiplication in the same expression. As a general rule all multiplications can be considered as matrix multiplication unless the dimensions are incompatible, in which case they are scalar multiplication. In particular a division sign denotes scalar multiplication. Thus for δ vector and n scalar $\delta/\sqrt{n}$ denotes scalar multiplication.

To avoid an orgy of indexes we shall also use matrix notation for differential operators.

The vector of partial derivatives (gradient) operator is written as

$$\frac{\partial}{\partial x}h(x) = \begin{pmatrix} \frac{\partial}{\partial x_1}h(x) \\ \frac{\partial}{\partial x_2}h(x) \\ \vdots \\ \frac{\partial}{\partial x_n}h(x) \end{pmatrix}.$$

The matrix of second partial derivatives (Hessian) operator is written as

$$\frac{\partial^2}{\partial x \partial x'}h(x) = \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1}h(x) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n}h(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1}h(x) & \cdots & \frac{\partial^2}{\partial x_n \partial x_n}h(x) \end{pmatrix}.$$

And finally the matrix of partial derivatives (Jacobi matrix) of a transformation $y(x)$ is written as

$$\frac{\partial y}{\partial x'} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}.$$

Note how the dimensions and indexes of the result formally match those of x in the gradient case, xx' in the Hessian case and yx' in the Jacobi case. The transposed gradient operator is correspondingly written as $\frac{\partial}{\partial x'}$.

By the integral of a matrix we shall (of course) mean the matrix consisting of the integral of each component.

By a notation like

$$\frac{\partial}{\partial \xi}f(y|\xi^0),$$

we shall mean the derivative of f with respect to ξ evaluated at the point ξ^0, thus avoiding the more cumbersome

$$\frac{\partial}{\partial \xi}f(y|\xi)\Big|_{\xi=\xi^0}.$$

In a further effort to economize notation, we shall allow the variable of a density (and other functions) to indicate the function in question. For example, $f(x)$ will denote the density of a variable X, while $f(y)$ denotes the (different) density of Y[1]. Thus we avoid having to "tag" a density name which really conveys no extra information to each variable introduced: Let X have the density f and Y have the density h and ... If there is any

[1]This is not strictly formal, of course, but common mathematical notation seldom is, anyway. In my view, allowing a slightly informal structure is actually an important feature of mathematical language intended for humans. A strictly formal language would be something like pure first order predicate logic, which would probably make this Cand. Scient. thesis require more pages than Encyclopedia Britannica.

chance of confusion the relevant variable will be clarified by a foot-script: $f_X(x)$.

The variance-covariance matrix of a random vector will be denoted by "Var", giving the ordinary variance in the one-dimensional case. The covariance matrix of two (different) random vectors will be denoted by "Cov".

In some proofs we shall make use of the "square root" of a matrix. The square root of a matrix A is denoted by $A^{1/2}$. We shall never actually have to determine any square root matrices. (It can be done through a spectral decomposition.) All we need to know is that for A positive definite, the square root matrix exists, is positive definite and has the property $A^{1/2}A^{1/2} = A$.

The multi-normal distribution is denoted by N. The dimension is determined by the parameters and will not be explicitly stated. However, round brackets will usually enclose the parameters in the one-dimensional case, while curly brackets will be used elsewhere.

Chapter 2

Maximum likelihood estimators

The aim of the present chapter is to study the wide and narrow ML-estimator in the framework given in the introduction. The performance criterion will be the limiting squared error risk function:

Definition: Let $\hat{\mu}$ be an estimator of μ, and denote by L the limit variable of $\sqrt{n}(\hat{\mu}-\mu)$. The squared error risk function is then defined by

$$r(\theta,\delta) = EL^2.$$

Note that the risk function will depend on the particular point of the narrow model, θ, and the "normed deviation" from the narrow model, δ. This is of course a consequence of the fact that the limit distribution of $\sqrt{n}(\hat{\mu}-\mu)$ will depend on these parameters. (Remember that in our large sample framework δ is held fixed while γ varies so that $\delta = \sqrt{n}(\gamma-\gamma^0)$.)

Remark 1: An alternative definition would be to define r as the limit of $nE(\hat{\mu}-\mu)^2$. Usually the two definitions will agree. However, the first definition corresponds to the procedure we shall actually use to derive the limits. So by choosing that definition we will not have to worry about this minor technical point. (Confer Lehmann (1983) p. 341.)

Remark 2: When actually computing numerical values we shall use the square root of the risk function. This will give the risk in the same units as the estimand and should be more meaningful for most practical purposes. (This is analogous to the question of variance versus standard deviation as a measure of dispersion.)

It will also prove convenient to have a risk concept for vector-valued estimators, so we define analogously:

Definition: Let $\hat{\xi}$ be an estimator of ξ, and denote by L the limit variable of $\sqrt{n}(\hat{\xi}-\xi)$. The squared error risk matrix is then defined by

$$R(\theta,\delta) = ELL'.$$

In section 2.7 we shall investigate the "quality" of the large sample risk functions as defined above, so we also define the following exact risk function for finite n.

Definition: Let $\hat{\mu}$ be an estimator of μ based on the n first observations. The exact risk function is defined by

$$r_n(\theta, \delta) = nE(\hat{\mu} - \mu)^2.$$

The exact risk function is given in terms of δ to facilitate comparison with the large sample risk function. In order to evaluate the exact risk function at a given parameter point $(\theta', \gamma')'$, substitute as usual $\delta = \sqrt{n}(\gamma - \gamma^0)$. A large sample approximation to the exact risk function at the given parameter point is correspondingly $r\left(\theta, \sqrt{n}(\gamma - \gamma^0)\right)$. The approximation should be good when n is large and $(\gamma - \gamma^0)$ is small.

The first sections of this chapter will be of a quite technical nature. If the reader is not interested in these technical details, he could skip Sections 2.1–2.3, read quickly through the results of Sections 2.4 and 2.5 and then start reading from section 2.6

2.1 Assumptions

We shall now state a set of (mild) regularity conditions on the density $f(y|\xi)$ which are sufficient to guarantee all required convergences in Chapters 2 and 3.

A1. The parameter space is a subset of Euclidean r-dimensional space with the "null" point ξ^0 as an inner point.

A2. The set $\{y : f(y|\xi) > 0\}$ does not depend on ξ.

A3. The density $f(y|\xi)$ has continuous partial derivatives with respect to ξ of order 3 for almost all y in a neighbourhood of ξ^0.

A4. The integral $\int f(y|\xi)\, dy$ can be differentiated twice under the integral sign with respect to ξ at the point ξ^0.

A5. The matrix $-E\frac{\partial^2}{\partial\xi\partial\xi'}\log f(Y|\xi^0)$ is positive definite.

A6. With probability tending to 1, the ML-estimators $\hat{\xi}_{\text{wide}}$ and $\hat{\theta}_{\text{narr}}$ are obtained as the unique solutions to the corresponding likelihood equations: $\frac{\partial}{\partial\xi}l(\xi) = 0$ and $\frac{\partial}{\partial\theta}l(\theta, \gamma_0) = 0$.

A7. There exists a function $g(y, \theta)$ dominating $f(y|\theta, \gamma)$ for γ in a neighbourhood of γ^0 such that the following integrals are finite:

$$\int \frac{\partial}{\partial\xi_j}\log f(y|\xi^0)g(y, \theta)\, dy,$$

$$\int \frac{\partial}{\partial \xi_j} \log f(y|\xi^0) \frac{\partial}{\partial \xi_k} \log f(y|\xi^0) \frac{\partial}{\partial \xi_l} \log f(y|\xi^0) g(y,\theta)\, dy,$$

$$\int \frac{\partial^2}{\partial \xi_j \partial \xi_k} \log f(y|\xi^0) g(y,\theta)\, dy.$$

A8. There exist functions $m_{jkl}(y)$ dominating $\frac{\partial^3}{\partial \xi_j \partial \xi_k \partial \xi_l} \log f(y|\xi)$ for ξ in a neighbourhood of ξ^0 such that (with g defined as above)

$$\int m_{jkl}(y) g(y,\theta)\, dy$$

is finite.

A9. There exist functions $g_j(y,\theta)$ dominating $\frac{\partial}{\partial \gamma_j} f(y|\theta,\gamma)$ for γ in a neighbourhood of γ^0 such that

$$\int \frac{\partial}{\partial \xi_k} \log f(y|\xi^0) \frac{\partial}{\partial \xi_l} \log f(y|\xi^0) g_j(y,\theta)\, dy$$

is finite.

A10. There exist functions $g_{jk}(y,\theta)$ dominating $\frac{\partial^2}{\partial \gamma_j \partial \gamma_k} f(y|\theta,\gamma)$ for γ in a neighbourhood of γ^0 such that

$$\int \frac{\partial}{\partial \xi_l} \log f(y|\xi^0) g_{jk}(y,\theta)\, dy$$

is finite.

A11. $\mu(\xi)$ has continuous partial derivatives of order two in a neighbourhood of ξ^0. (Note that continuity also implies boundedness of the second partial derivatives in a neighbourhood of ξ^0.)

Assumptions A1–A6 and A8 are with small variations the conditions given by Lehmann (1983) to assure the well-known limit distribution of the ML-estimator in the standard i.i.d. case. A7, A9 and A10 are similar in spirit to A8 and should be fulfilled in most cases. A11 requires regularity of the estimand under study μ, and should probably be satisfied for almost all interesting estimands.

2.2 Some convergence lemmas

Before embarking on the main results, we need to prove a few preliminary lemmas, which is the purpose of this section.

Lemma 2.2.1 *Let h be a real-valued function, and suppose $Eh(Y_{ni}) \to Eh(Y)$. Then*

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(Y_{ni}) \xrightarrow{P} Eh(Y).$$

Remark: It will clearly be sufficient for $f(y,\theta,\gamma)$ to be dominated by a function $g(y,\theta)$ for γ in a neighbourhood of γ^0 such that

$$\int h(y)g(y,\theta)\,dy$$

is finite. For in that case Lebesgue's dominated convergence theorem can be invoked to guarantee

$$\int h(y)f(y,\theta,\gamma^0+\delta/\sqrt{n})\,dy \to \int h(y)f(y,\theta,\gamma^0)\,dy,$$

which is the assumption of the lemma. A similar remark can also be applied to the following two lemmas.

Proof: Let $\mu_n = Eh(Y_{ni}), \mu = Eh(Y)$ and denote by $\Phi_n(t)$ the characteristic function of $h(Y_{ni})$. According to Hjort (1980) p. 95 we may write

$$\Phi_n(t) = 1 + i\mu_n t + o(t).$$

It follows that the characteristic function of $\bar{h}$ is given by

$$\cdot \left(1 + i\mu_n t/n + o(t/n)\right)^n =$$

$$\left(1 + i\mu t/n + i\frac{(\mu_n-\mu)t}{n} + o(t/n)\right)^n =$$

$$\left(1 + i\mu t/n + o(t/n)\right)^n \to e^{i\mu t}.$$

(Confer Hjort's compendium.) The conclusion now follows since $e^{i\mu t}$ is the characteristic function for μ. □

Lemma 2.2.2 *Suppose $\hat{\xi} \xrightarrow{P} \xi^0$. Let $h(y,\xi)$ be a real-valued function, such that*

$$Eh(Y_{ni},\xi^0) \to Eh(Y,\xi^0), \tag{2.1}$$

and suppose there exist functions $g_j(y)$ dominating $\frac{\partial}{\partial\xi_j}h(y,\xi)$ for ξ in a neighbourhood of ξ^0 with

$$Eg_j(Y_{ni}) \to Eg_j(Y) < \infty. \tag{2.2}$$

Then

$$\frac{1}{n}\sum_{i=1}^{n} h(Y_{ni},\hat{\xi}) \xrightarrow{P} Eh(Y,\xi^0).$$

Proof: Use a first order Taylor expansion to obtain:

$$\frac{1}{n}\sum_{i=1}^{n} h(Y_{ni},\hat{\xi}) = \frac{1}{n}\sum_{i=1}^{n} h(Y_{ni},\xi^0) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \xi'} h(Y_{ni},\tilde{\xi})(\hat{\xi}-\xi^0)$$

where $\tilde{\xi}$ is a point on the line between $\hat{\xi}$ and ξ^0 and thus converges in probability to ξ^0 since $\hat{\xi}$ does. From Lemma 2.2.1 and assumption (2.1) we see that the first term on the right hand side converges in probability to $Eh(Y,\xi^0)$. Thus the conclusion follows as soon as we have demonstrated that the second term converges to zero. Since $(\hat{\xi}-\xi^0)$ converges to zero, it is sufficient to demonstrate that $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \xi'} h(Y_{ni},\tilde{\xi})$ is bounded in probability. Componentwise we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \xi_j} h(Y_{ni},\tilde{\xi})\right| \leq \frac{1}{n}\sum_{i=1}^{n} \left|\frac{\partial}{\partial \xi_j} h(Y_{ni},\tilde{\xi})\right| \leq \frac{1}{n}\sum_{i=1}^{n} g_j(Y_{ni}),$$

the last inequality being valid with probability tending to one. This bound converges in probability to $Eg_j(Y)$ due to Lemma 2.2.1 and assumption (2.2), which completes the proof. □

Lemma 2.2.3 *Let h be a vector-valued function satisfying the following for some (finite) μ and K:*

$$E\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(Y_{ni}) \to \mu, \tag{2.3}$$

$$\mathrm{Var}\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(Y_{ni}) \to K. \tag{2.4}$$

Suppose further that the third order moments of $h(Y_{ni})$ are bounded by constants independent of n:

$$E|h_j(Y_{ni})h_k(Y_{ni})h_l(Y_{ni})| \leq m_{jkl}. \tag{2.5}$$

Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(Y_{ni}) \xrightarrow{D} N\{\mu,K\}.$$

Proof: Let $\mu_n = Eh(Y_{ni})$ and $K_n = \mathrm{Var}\, h(Y_{ni})$. By the Cramér–Wold theorem the conclusion of the lemma is equivalent to

$$a' \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(Y_{ni}) \xrightarrow{D} N(a'\mu, a'Ka)$$

being valid for all a. By (2.3) and (2.4) and the Cramér rules this is equivalent to

$$\sum_{i=1}^{n} \frac{1}{\sqrt{n}} (a'h(Y_{ni}) - a'\mu_n)(a'K_n a)^{-1/2} \xrightarrow{D} N(0,1).$$

This is a sum of independent variables with expectation 0 and sum of variances 1. From the Lindeberg–Lyapunov theorem we obtain a sufficient condition for the required convergence. The sum of the absolute value of the variables third order moments should converge to zero:

$$\sum_{i=1}^{n} E\left(\frac{1}{\sqrt{n}}\right)^3 |a'h(Y_{ni}) - a'\mu_n|^3 (a'K_n a)^{-3/2} \to 0 \quad \Leftrightarrow$$

$$\frac{1}{\sqrt{n}} E\,|a'h(Y_{ni}) - a'\mu_n|^3 \to 0.$$

And this convergence is implied by (2.5). This concludes the proof of the lemma. □

Remark: The assumptions of the above lemma could be weakened somewhat. We have nevertheless stated the lemma in this form since the given assumptions are usually easy to verify and will be satisfied in all situations of interest to us.

Lemma 2.2.4 *Let x_n be the solution of*

$$A_n x_n = b_n,$$

and suppose $b_n \xrightarrow{D} b$ and $A_n \xrightarrow{P} A$ where A is invertible. Then

$$x_n \xrightarrow{D} A^{-1}b.$$

Proof: With probability converging to 1, A_n is invertible since A is. (Consider for example the determinant of A_n which by Slutsky will converge to the determinant of A, and hence be different from zero with probability converging to 1.) Therefore $x_n = A_n^{-1} b_n$ with probability tending to 1, and the rest follows from Cramér and Slutsky.

2.3 The information matrix J

The information matrix of the wide model will play a central part in what follows, and we devote this section to give a few useful results.

We define our information matrix as the information matrix of the wide model but evaluated at the narrow model only.

$$J = J(\theta) = \mathrm{Var}\,\frac{\partial}{\partial \xi} \log f(Y|\xi^0).$$

We shall need a few alternative forms which we state as a small lemma.

Lemma 2.3.1 *The information matrix can be written*

$$J = E\frac{\partial}{\partial \xi} \log f(Y|\xi^0)\frac{\partial}{\partial \xi'} \log f(Y|\xi^0) = -E\frac{\partial^2}{\partial \xi \partial \xi'} \log f(Y|\xi^0).$$

Proof: This is a standard result, but we shall give a proof anyway, mainly in order to introduce the reader to our quite compact notation for differential operators.

Consider first

$$\begin{aligned} E\frac{\partial}{\partial \xi} \log f(Y|\xi^0) &= \int \frac{\frac{\partial}{\partial \xi} f(y|\xi^0)}{f(y|\xi^0)} f(y|\xi^0)\, dy \\ &= \frac{\partial}{\partial \xi} \int f(y|\xi^0)\, dy \\ &= 0, \end{aligned}$$

since we have assumed differentiability under the integral sign. This proves the first equality of the lemma. To prove the other consider

$$\begin{aligned} -E\frac{\partial^2}{\partial \xi \partial \xi'} \log f(Y|\xi^0) &= -E\frac{\partial}{\partial \xi}\frac{\frac{\partial}{\partial \xi'} f(Y|\xi^0)}{f(Y|\xi^0)} \\ &= -E\frac{f(Y|\xi^0)\frac{\partial^2}{\partial \xi \partial \xi'} f(Y|\xi^0) - \frac{\partial}{\partial \xi} f(Y|\xi^0)\frac{\partial}{\partial \xi'} f(Y|\xi^0)}{f^2(Y,\xi)}. \end{aligned}$$

Again as a consequence of differentiability under the integral sign, the first term on the right-hand side vanishes and we are left with

$$E\frac{\partial}{\partial \xi} \log f(Y|\xi^0)\frac{\partial}{\partial \xi'} \log f(Y|\xi^0).$$

This concludes the proof. If the reader does not like our vector generalization of the quotient rule for differentiation, he could equally well verify the equalities component-wise. □

Now introduce a partition of J and its inverse:

$$J = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix}, \quad J^{-1} = \begin{pmatrix} J^{11} & J^{12} \\ J^{21} & J^{22} \end{pmatrix}.$$

The dimensions of the sub-matrices correspond to those of the parameter vectors θ and γ. ($J_{11}^{p\times p}$, $J_{22}^{q\times q}$ etc.)

In the following sections we shall need a few correspondences between the sub-matrices of J and J^{-1}, which can be obtained by observing that

$$\begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix} \begin{pmatrix} J^{11} & J^{12} \\ J^{21} & J^{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

Multiplying out the product on the left-hand side and equating components give four equations:

$$J_{11}J^{11} + J_{12}J^{21} = I, \tag{2.6}$$

$$J_{11}J^{12} + J_{12}J^{22} = 0, \tag{2.7}$$

$$J_{21}J^{11} + J_{22}J^{21} = 0, \tag{2.8}$$

$$J_{21}J^{12} + J_{22}J^{22} = I. \tag{2.9}$$

From these we easily obtain two more equations which will prove useful.

$$J^{12}(J^{22})^{-1} = -J_{11}^{-1}J_{12}, \tag{2.10}$$

$$J^{11} = J_{11}^{-1} + J_{11}^{-1}J_{12}J^{22}J_{21}J_{11}^{-1}. \tag{2.11}$$

2.4 Limiting risk for the ML-estimators

In the present section, we shall derive the risk matrices of the two ML-estimators for ξ, $\hat{\xi}_{\text{wide}}$ based on wide model estimation and $\hat{\xi}_{\text{narr}}$ based on narrow model estimation. These risk matrices will in turn be used to determine the risk function for the corresponding estimators for μ. But first of all we shall prove the following lemma:

Lemma 2.4.1 *Both the narrow and wide estimator of ξ are consistent for ξ^0:*

$$\hat{\xi}_{\text{wide}} \xrightarrow{P} \xi^0,$$

$$\hat{\xi}_{\text{narr}} \xrightarrow{P} \xi^0.$$

(We consider γ^0 to be the narrow estimator of γ.)

Proof: This is a standard result under null model conditions ($\delta = 0$), and a proof can be found in Lehmann (1983) p. 430. We shall use a modified version of Lehmann's proof, adapted to our local neighbourhood framework.

We shall prove the result first for the wide model estimator. The starting point is a third order Taylor expansion of the log-likelihood function:

$$\begin{aligned} l(\xi) &= l(\xi^0) + \frac{\partial}{\partial \xi'} l(\xi^0)(\xi - \xi^0) + \frac{1}{2}(\xi - \xi^0)' \frac{\partial^2}{\partial \xi \partial \xi'} l(\xi^0)(\xi - \xi^0) + \\ & \frac{1}{6} \sum_{j,k,l} \frac{\partial^3}{\partial \xi_j \partial \xi_k \partial \xi_l} l(\tilde{\xi})(\xi_j - \xi_j^0)(\xi_k - \xi_k^0)(\xi_l - \xi_l^0). \end{aligned}$$

Divide by n and rearrange:

$$\begin{aligned} \frac{1}{n}\big(l(\xi) - l(\xi^0)\big) &= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \xi'} \log f(Y_{ni}|\xi^0)(\xi - \xi^0) + \\ & \frac{1}{2}(\xi - \xi^0)' \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \xi \partial \xi'} \log f(Y_{ni}|\xi^0)(\xi - \xi^0) + \\ & \frac{1}{6} \sum_{j,k,l} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3}{\partial \xi_j \partial \xi_k \partial \xi_l} \log f(Y_{ni}|\tilde{\xi})(\xi_j - \xi_j^0)(\xi_k - \xi_k^0)(\xi_l - \xi_l^0) \\ &= S_1 + S_2 + S_3. \end{aligned}$$

Suppose that ξ is located on the surface of an r-ball around ξ^0 with small radius a. If we can demonstrate that the maximum of the right-hand side of the above equation is negative with probability tending to 1 for all sufficiently small a, the conclusion of the lemma follows. For in this case the log-likelihood function must have a local maximum inside the r-ball, again with probability tending to 1. And by our assumptions this maximum must be the unique MLE.

Consider S_1 first. By Lemma 2.2.1 and assumption A7 we have:

$$S_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \xi'} \log f(Y_{ni}|\xi^0)(\xi - \xi^0) \xrightarrow{P} E \frac{\partial}{\partial \xi'} \log f(Y|\xi^0)(\xi - \xi^0) = 0.$$

The sum S_2 may be written as:

$$S_2 = -\frac{1}{2}(\xi - \xi^0)'J(\xi - \xi^0) + \frac{1}{2}(\xi - \xi^0)'\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\xi\partial\xi'}\log f(Y_{ni}|\xi^0) + J\right)(\xi - \xi^0).$$

Since J is positive definite the maximum value of the first term is given by $-\frac{1}{2}\lambda_{\min}a^2$, where $\lambda_{\min}$ is the smallest eigenvalue of J. The second term converges to

$$\frac{1}{2}(\xi - \xi^0)'\left(E\frac{\partial^2}{\partial\xi\partial\xi'}\log f(Y|\xi^0) + J\right)(\xi - \xi^0) = 0$$

by Lemma 2.2.1 and assumption A7.

Now finally consider S_3. Note that $\tilde{\xi}$ is a point on the line between ξ and ξ^0 and thus has distance less than a to ξ^0. Therefore we may conclude from assumption A8 that, for sufficiently small a, there exist dominating functions $m_{jkl}(y)$ such that:

$$|S_3| \leq \frac{1}{6}\sum_{j,k,l}\frac{1}{n}\sum_{i=1}^{n}|\frac{\partial^3}{\partial\xi_j\partial\xi_k\partial\xi_l}\log f(Y_{ni}|\tilde{\xi})|a^3 \leq \frac{1}{6}\sum_{j,k,l}\frac{1}{n}\sum_{i=1}^{n}m_{jkl}(Y_{ni})a^3.$$

By the same assumption and Lemma 2.2.1 this bound converges in probability to

$$\frac{1}{6}\sum_{j,k,l}Em_{jkl}(Y)a^3.$$

And this is smaller in magnitude than $\frac{1}{2}\lambda_{\min}a^2$ for sufficiently small a.

Combining all this information we see that for all sufficiently small a, $S_1 + S_2 + S_3$ is negative with probability tending to 1. This concludes the proof for the wide case.

The narrow model estimator $\hat{\theta}_{\text{narr}}$ is by definition obtained by maximizing the likelihood function $l(\theta, \gamma^0)$ with respect to θ, as opposed to the wide model estimator which is the maximum with respect to both θ and γ. The proof in the narrow case will, however, be completely analogous to the wide case. Simply replace Taylor expansion with respect to ξ by Taylor expansion with respect to θ. □

We are now in a position to derive the limit distribution for the two competing estimators of ξ:

Lemma 2.4.2 *The limit distribution of the wide model estimator of ξ is given by*

$$\sqrt{n}(\hat{\xi}_{\text{wide}} - \xi) \xrightarrow{D} N\{0, J^{-1}\}.$$

This corresponds to the risk matrix

$$R_{\text{wide}}(\theta, \delta) = J^{-1}.$$

Proof: By the defining property of the MLE

$$\sum_{i=1}^{n} \frac{\partial}{\partial \xi} \log f(Y_{ni}|\hat{\xi}_{\text{wide}}) = 0.$$

A Taylor expansion around ξ^0 yields

$$\sum_{i=1}^{n} \frac{\partial}{\partial \xi} \log f(Y_{ni}|\xi^0) + \sum_{i=1}^{n} \frac{\partial^2}{\partial \xi \partial \xi'} \log f(Y_{ni}|\tilde{\xi})(\hat{\xi}_{\text{wide}} - \xi^0) = 0$$

(Actually we use a separate Taylor expansion for each component of the original vector, which means that $\tilde{\xi}$ in the matrix $\sum_{i=1}^{n} \frac{\partial^2}{\partial \xi \partial \xi'} \log f(Y_{ni}|\tilde{\xi})$ really will be different in each line. This will be seen to be of no consequence, however, so we stick to our compact, but somewhat imprecise notation.)

A little rearranging gives:

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \xi \partial \xi'} \log f(Y_{ni}|\tilde{\xi}) \sqrt{n}(\hat{\xi}_{\text{wide}} - \xi^0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \xi} \log f(Y_{ni}|\xi^0),$$

which may be written

$$A_{n,\text{wide}} \sqrt{n}(\hat{\xi}_{\text{wide}} - \xi^0) = b_{n,\text{wide}}.$$

Consider $A_{n,\text{wide}}$ first. We use Lemma 2.2.2 componentwise to obtain the limit. For each component we know that $\tilde{\xi}$ converges in probability to ξ^0 since $\hat{\xi}$ does. This, together with assumptions A7 and A8 secure the result:

$$A_{n,\text{wide}} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \xi \partial \xi'} \log f(Y_{ni}|\tilde{\xi}) \xrightarrow{P} -E \frac{\partial^2}{\partial \xi \partial \xi'} \log f(Y|\xi^0) = J.$$

Now consider $b_{n,\text{wide}}$. If we can determine the limits of $Eb_{n,\text{wide}}$ and $\text{Var}\, b_{n,\text{wide}}$, assumption A7 allows us to use Lemma 2.2.3. For the expectation we have:

$$\begin{aligned} Eb_{n,\text{wide}} &= E \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \xi} \log f(Y_{ni}|\xi^0) \\ &= \sqrt{n} E \frac{\partial}{\partial \xi} \log f(Y_{n1}, \xi^0) \\ &= \int \sqrt{n} \frac{\partial}{\partial \xi} \log f(y|\xi^0) f(y|\theta, \gamma^0 + \delta/\sqrt{n})\, dy. \end{aligned}$$

Now Taylor-expand f with respect to γ inside the integral:

$$\int \sqrt{n}\frac{\partial}{\partial\xi}\log f(y|\xi^0)\left[f(y|\theta,\gamma^0)+\frac{1}{\sqrt{n}}\frac{\partial}{\partial\gamma'}f(y|\theta,\gamma^0)\delta+\frac{1}{n}\delta'\frac{\partial^2}{\partial\gamma\partial\gamma'}f(y|\theta,\tilde{\gamma}(y))\delta\right]dy.$$

The integral of the first term is zero since we have assumed differentiability under the integral sign. For the remainder term we know that $\tilde{\gamma}(y)$ is somewhere between γ^0 and $\gamma^0+\delta/\sqrt{n}$. Thus for sufficiently large n there exist, by assumption A10, dominating functions $g_{jk}(y,\theta)$ giving:

$$\left|\int\frac{1}{\sqrt{n}}\frac{\partial}{\partial\xi_l}\log f(y|\xi^0)\delta'\frac{\partial^2}{\partial\gamma\partial\gamma'}f(y|\theta,\tilde{\gamma}(y))\delta\,dy\right|\leq$$

$$\frac{1}{\sqrt{n}}\int\left|\frac{\partial}{\partial\xi_l}\log f(y|\xi^0)\right|\sum_{j,k}\left|\delta_j\frac{\partial^2}{\partial\gamma_j\partial\gamma_k}f(y|\theta,\tilde{\gamma}(y))\delta_k\right|dy\leq$$

$$\frac{1}{\sqrt{n}}\int\left|\frac{\partial}{\partial\xi_l}\log f(y|\xi^0)\right|\sum_{j,k}\left|\delta_j g_{jk}(y,\theta)\delta_k\right|dy\to 0.$$

The limit of $Eb_{n,\text{wide}}$ is consequently given by the limit of the second term in the Taylor expansion which is

$$\int\frac{\partial}{\partial\xi}\log f(y|\xi^0)\frac{\partial}{\partial\gamma'}\log f(y|\xi^0)\delta f(y|\xi^0)\,dy=\begin{pmatrix}J_{12}\\J_{22}\end{pmatrix}\delta.$$

Now for the variance of $b_{n,\text{wide}}$:

$$\begin{aligned}\operatorname{Var}b_{n,\text{wide}} &= \operatorname{Var}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\xi}\log f(Y_{ni}|\xi^0)\\ &= E\frac{\partial}{\partial\xi}\log f(Y_{ni}|\xi^0)\frac{\partial}{\partial\xi'}\log f(Y_{ni}|\xi^0)-E\frac{\partial}{\partial\xi}\log f(Y_{ni}|\xi^0)E\frac{\partial}{\partial\xi'}\log f(Y_{ni}|\xi^0).\end{aligned}$$

The second term is seen to converge to zero from the result above, so we can concentrate on the first term. Once more we Taylor-expand f to obtain:

$$E\frac{\partial}{\partial\xi}\log f(Y_{ni}|\xi^0)\frac{\partial}{\partial\xi'}\log f(Y_{ni}|\xi^0)=$$

$$\int\frac{\partial}{\partial\xi}\log f(y|\xi^0)\frac{\partial}{\partial\xi'}\log f(y|\xi^0)\left[f(y|\theta,\gamma^0)+\frac{1}{\sqrt{n}}\frac{\partial}{\partial\gamma'}f(y|\theta,\tilde{\gamma}(y))\delta\right]dy.$$

As before the remainder term is seen to converge to zero, this time due to the dominating functions from assumption A9. And the first term is nothing but J.

Consequently we have demonstrated that

$$b_{n,\text{wide}} \xrightarrow{D} N\{\begin{pmatrix} J_{12} \\ J_{22} \end{pmatrix} \delta, J\}.$$

From Lemma 2.2.4 we can now conclude that

$$\sqrt{n}(\hat{\xi}_{\text{wide}} - \xi^0) \xrightarrow{D} J^{-1}N\{\begin{pmatrix} J_{12} \\ J_{22} \end{pmatrix} \delta, J\} = N\{\begin{pmatrix} 0 \\ \delta \end{pmatrix}, J^{-1}\}.$$

Remembering that $\xi = \xi^0 + (0', \delta'/\sqrt{n})'$ this is equivalent to

$$\sqrt{n}(\hat{\xi}_{\text{wide}} - \xi) \xrightarrow{D} N\{0, J^{-1}\}.$$

□

We now turn to the narrow model estimator. The result is:

Lemma 2.4.3 *The limit distribution of the narrow model estimator of ξ is given by: (We consider the narrow estimate of γ to be γ_0.)*

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_{\text{narr}} - \theta \\ \gamma_0 - \gamma \end{pmatrix} \xrightarrow{D} N\{\begin{pmatrix} J_{11}^{-1}J_{12}\delta \\ -\delta \end{pmatrix}, \begin{pmatrix} J_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix}\},$$

which corresponds to the risk matrix:

$$R_{\text{narr}}(\theta, \delta) = \begin{pmatrix} J_{11}^{-1} + J_{11}^{-1}J_{12}\delta\delta'J_{21}J_{11}^{-1} & J_{11}^{-1}J_{12}\delta\delta' \\ \delta\delta'J_{21}J_{11}^{-1} & \delta\delta' \end{pmatrix}.$$

Proof: The proof is quite similar to the wide case. The narrow ML-estimator satisfies:

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y_{ni}|\hat{\theta}_{\text{narr}}, \gamma^0) = 0.$$

Taylor expansion, now with respect to θ, yields:

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y_{ni}|\xi^0) + \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \log f(Y_{ni}|\tilde{\theta}, \gamma^0)(\hat{\theta}_{\text{narr}} - \theta) = 0$$

Rearranging this expression we obtain

$$-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \log f(Y_{ni}|\tilde{\theta}, \gamma^0)\sqrt{n}(\hat{\theta}_{\text{narr}} - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y_{ni}|\xi^0),$$

which we write

$$A_{n,\text{narr}}\sqrt{n}(\hat{\theta}_{\text{narr}} - \theta) = b_{n,\text{narr}}.$$

This time $A_{n,\text{narr}}$ converges to J_{11}, and as before Lemma 2.2.2 together with assumptions A7 and A8 give the result.

The vector $b_{n,\text{narr}}$ is simply a sub-vector of $b_{n,\text{wide}}$ from the proof of the preceding lemma. The limit variable of $b_{n,\text{narr}}$ will thus be the corresponding sub-vector of the limit variable of $b_{n,\text{wide}}$. That is

$$b_{n,\text{narr}} \xrightarrow{D} N\{J_{12}\delta, J_{11}\}.$$

Having obtained the limits of $A_{n,\text{narr}}$ and $b_{n,\text{narr}}$, we may use Lemma 2.2.4 to conclude:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} J_{11}^{-1}N\{J_{12}\delta, J_{11}\} = N\{J_{11}^{-1}J_{12}\delta, J_{11}^{-1}\},$$

which immediately gives the conclusion of the lemma. □

Remark: The choice of the factor $\sqrt{n}$ in $\gamma = \gamma^0 + \delta/\sqrt{n}$ might have seemed quite arbitrary. But examination of the proofs of the preceding two lemmas show that a factor tending faster to infinity than $\sqrt{n}$ would make the narrow estimator better than the wide for all δ. And a factor tending slower to infinity than $\sqrt{n}$ would make the risk of the narrow estimator infinite for all $\delta \neq 0$. Thus we conclude that $\sqrt{n}$ is the right norming factor in order to obtain interesting large sample approximations.

In the special case of a one-dimensional γ the two preceding lemmas correspond to the proposition on p. 6 of EMMM.

Our next step is to derive the risk functions for the μ estimators. The following lemma will show how:

Lemma 2.4.4 *Let $\hat{\xi}$ be any estimator for ξ with existing risk matrix $R(\theta, \delta)$. The corresponding estimator for μ, $\hat{\mu} = \mu(\hat{\xi})$, has risk function given by:*

$$r(\theta, \delta) = \frac{\partial \mu}{\partial \xi'} R(\theta, \delta) \frac{\partial \mu}{\partial \xi},$$

where the partial derivatives are computed at the null point ξ^0.

Further, if the ξ estimator has a normal limit distribution, then the μ estimator also has a normal limit distribution:

$$\sqrt{n}(\hat{\xi} - \xi) \xrightarrow{D} N\{\eta, \Sigma\} \Rightarrow \sqrt{n}(\hat{\mu} - \mu) \xrightarrow{D} N\left(\frac{\partial \mu}{\partial \xi'}\eta, \frac{\partial \mu}{\partial \xi'}\Sigma\frac{\partial \mu}{\partial \xi}\right).$$

Proof: Taylor-expand both terms in $\sqrt{n}(\hat{\mu}-\mu)$ around ξ^0:

$$\sqrt{n}\big(\mu(\hat{\xi})-\mu(\xi)\big) = \sqrt{n}\Big[\,\mu(\xi^0)+\frac{\partial}{\partial\xi'}\mu(\xi^0)(\hat{\xi}-\xi^0)+\frac{1}{2}(\hat{\xi}-\xi^0)'\frac{\partial^2}{\partial\xi\partial\xi'}\mu(\tilde{\xi})(\hat{\xi}-\xi^0)$$
$$-\mu(\xi^0)-\frac{\partial}{\partial\xi'}\mu(\xi^0)(\xi-\xi^0)-\frac{1}{2}(\xi-\xi^0)'\frac{\partial^2}{\partial\xi\partial\xi'}\mu(\tilde{\tilde{\xi}})(\xi-\xi^0)\Big].$$

We shall prove that the two remainder terms go to zero which implies

$$\sqrt{n}(\hat{\mu}-\mu) \doteq \frac{\partial}{\partial\xi'}\mu(\xi^0)\sqrt{n}(\hat{\xi}-\xi),$$

from which the conclusion of the lemma follows immediately.

Now consider the first remainder:

$$\sqrt{n}\frac{1}{2}(\hat{\xi}-\xi^0)'\frac{\partial^2}{\partial\xi\partial\xi'}\mu(\tilde{\xi})(\hat{\xi}-\xi^0).$$

Observe that since $\sqrt{n}(\hat{\xi}-\xi)$ has a limit distribution, $\sqrt{n}(\hat{\xi}-\xi^0)$ also has a limit distribution and $(\hat{\xi}-\xi^0)$ converges to zero in probability. Further $\frac{\partial^2}{\partial\xi\partial\xi'}\mu(\tilde{\xi})$ may be dominated by a constant matrix by assumption A11, with probability tending to one. From these facts and the Cramér rules follows that the remainder converges in probability to zero.

The second remainder is easily seen to converge to zero, remembering that

$$(\xi-\xi^0) = \begin{pmatrix} 0 \\ \delta/\sqrt{n} \end{pmatrix}.$$

□

The main results of this section are now obtained as an immediate consequences of the preceding lemmas.

Theorem 2.4.1 *The wide estimator of μ has risk function given by*

$$r_{\text{wide}}(\theta,\delta) = \frac{\partial\mu}{\partial\xi'}J^{-1}\frac{\partial\mu}{\partial\xi},$$

which can be written as

$$r_{\text{wide}}(\theta,\delta) = b'J^{22}b+\tau_0^2,$$

where

$$b = b(\theta) = J_{21}J_{11}^{-1}\frac{\partial\mu}{\partial\theta}-\frac{\partial\mu}{\partial\gamma} \qquad \textit{and} \qquad \tau_0^2 = \tau_0^2(\theta) = \frac{\partial\mu}{\partial\theta'}J_{11}^{-1}\frac{\partial\mu}{\partial\theta}.$$

The limit distribution is

$$\sqrt{n}(\hat{\mu}_{\text{wide}} - \mu) \xrightarrow{D} N(0, \frac{\partial \mu}{\partial \xi'} J^{-1} \frac{\partial \mu}{\partial \xi}).$$

Proof: The only thing left to prove is the equivalence between the two expressions for $r_{\text{wide}}(\theta, \delta)$. The first expression is

$$\left(\frac{\partial \mu}{\partial \theta'}, \frac{\partial \mu}{\partial \gamma'}\right) \begin{pmatrix} J^{11} & J^{12} \\ J^{21} & J^{22} \end{pmatrix} \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}.$$

Substitute $J^{12} = -J_{11}^{-1} J_{12} J^{22}$, $J^{21} = -J^{22} J_{21} J_{11}^{-1}$ and $J^{11} = J_{11}^{-1} + J_{11}^{-1} J_{12} J^{22} J_{21} J_{11}^{-1}$ from (2.10) and (2.11) on p. 16. The second expression for r_{wide} is now obtained by multiplying out the matrix product.

Theorem 2.4.2 *The narrow estimator of μ has risk function given by*

$$r_{\text{narr}}(\theta, \delta) = \frac{\partial \mu}{\partial \xi'} \begin{pmatrix} J_{11}^{-1} + J_{11}^{-1} J_{12} \delta\delta' J_{21} J_{11}^{-1} & J_{11}^{-1} J_{12} \delta\delta' \\ \delta\delta' J_{21} J_{11}^{-1} & \delta\delta' \end{pmatrix} \frac{\partial \mu}{\partial \xi},$$

which we write as (with b and τ_0^2 as in the preceding theorem):

$$r_{\text{narr}}(\theta, \delta) = b'\delta\delta'b + \tau_0^2.$$

The limit distribution is

$$\sqrt{n}(\hat{\mu}_{\text{narr}} - \mu) \xrightarrow{D} N(b'\delta, \tau_0^2).$$

2.5 Regression generalization

Suppose now that our data $Y_{n1}, Y_{n2}, \ldots, Y_{nn}$ are not identically distributed any longer, but have a regression structure: Given some x_i, Y_{ni} is distributed according to the density $f(y|\theta, \gamma, x_i)$, where $\gamma = \gamma^0 + \delta/\sqrt{n}$ as before. We could now generalize all our results by replacing the definition of J by

$$J = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E_0 \frac{\partial^2}{\partial \xi \partial \xi'} \log f(Y_{ni}|\xi^0, x_i)$$

where the subscript 0 signifies that the expectation is computed under the null distribution $f(y|\theta, \gamma^0, x_i)$. By imposing various regularity conditions on the x_i as well as on the density, it would be possible to ensure that all our results carry over verbatim. The regularity conditions should, roughly speaking, ensure that the x_i do not tend to extreme values. This should imply the convergence of the J matrix as defined above and normal limits

through Lindeberg's theorem.

We shall take a slightly different point of view, however, and in a sense comprise all the regularity conditions into one. The point is to consider the x_i's as random variables. This should be reasonable in most contexts where the x_i's are not decided by the experimenter. Even in cases where the experimenter is responsible for the x_i values, the random variable approach could be reasonable if the experimenter arranged the x_i's so as to comply with the assumption.

Our assumption is that $x_1, x_2, \ldots, x_n$ is independent and identically distributed with a common density $f(x|\zeta)$, possibly depending on a parameter ζ, but not on ξ. $f(y|\theta, \gamma, x)$ is then considered to be the conditional density for Y_{ni} given x_i. To be exact we assume that the simultaneous density of the n first x_i's and Y_{ni}'s is $\prod_{i=1}^{n} f(y_i|\xi, x_i) f(x_i|\zeta)$. In this manner the pairs $(Y'_{ni}, x'_i)'$ will be i.i.d. and the theory developed can be used unchanged. Simply replace Y_{ni} by $(Y'_{ni}, x'_i)'$. For an example of a similar device used in a different context, see Neuhaus (1985).

The first thing to verify is that our estimators remain unchanged from the fixed x_i approach. This is intuitively reasonable since the distribution of the x_i's does not depend on ξ. The formal verification consists simply of noting that the likelihood equations remain unchanged since

$$\frac{\partial}{\partial \xi} \log\Big\{ f(Y_{ni}|\xi, x_i) f(x_i|\zeta) \Big\} = \frac{\partial}{\partial \xi} \log f(Y_{ni}|\xi, x_i). \tag{2.12}$$

The second thing to note is that the J matrix will now be the expectation in the simultaneous distribution of Y and x, which we can write as (let now Y be distributed as $f(y|\xi^0, x)$):

$$\begin{aligned} J &= -E \frac{\partial^2}{\partial \xi \partial \xi'} \log\Big\{ f(Y|\xi^0, x) f(x|\zeta) \Big\} \\ &= -E \frac{\partial^2}{\partial \xi \partial \xi'} \log f(Y|\xi^0, x) \\ &= -EE \left[\frac{\partial^2}{\partial \xi \partial \xi'} \log f(Y|\xi^0, x) \Big| x \right] \\ &= EJ(x). \end{aligned}$$

where $J(x)$ is the "conditional information matrix" given x, and the last expectation is under the distribution of x alone.

Finally note that $r(\theta, \delta)$ will now be the limiting squared error risk function in the simultaneous distribution.

Remark 1: The alert reader may have discovered that although we have restored the i.i.d. assumption, there is one subtle change in the situation from the non-regression context. The parameter ξ is now only a subset of the parameters of the full model. Could this affect our results? The answer is no, as can be verified by the following simple argument.

If ζ were known, the structure of the model would be exactly as before with $(Y_{ni}', x_i')'$ in the place of Y_{ni}, and $f(y|\xi, x)f(x|\zeta)$ in the place $f(y|\xi)$. In that case we could freely use all our old risk formulas. (They would of course depend on ζ.) But (2.12) shows that the form of the estimator of ξ will be the same whether ζ is known or not. And in this case the value of the risk function, as a function of ζ, will of course not depend on whether ζ is actually known. Thus we can simply compute risk functions as if ζ were known, by our old formula, and they will stay valid also in the case of unknown ζ.

Remark 2: One could argue that it is artificial to consider the average loss over all values of x, since at the time of estimation it is known which values of x that occurred. An answer to this objection is to point out that this is not different from the situation concerning Y. Actually, all classical error measures are based on averages over data which are known not to occur. More will be said about this later, see p. 74.

Suppose now that one wanted to estimate J. One could do so without actually specifying the parametric structure of $f(x|\zeta)$. A natural estimate could then be obtained by replacing θ with $\hat{\theta}$ (either the narrow or the wide version) and the distribution of x by the empirical distribution function. This would lead to the estimate

$$\hat{J} = \frac{1}{n}\sum_{i=1}^{n} J(x_i, \hat{\theta}).$$

This estimator is consistent, see Neuhaus (1985) for a proof.

2.6 Comparison of risk for narrow and wide estimator

We are now in a position to give a precise criterion for when narrow estimation is better than wide estimation.

Theorem 2.6.1 *Narrow estimation is better than wide estimation for a given estimand μ if and only if*

$$b'\delta\delta'b < b'J^{22}b. \tag{2.13}$$

(If $b = 0$ the two estimators are equivalent for all δ. See Theorem 2.4.1 for the definition of b.)
Narrow estimation is better than wide estimation for all estimands with $b \neq 0$ if

$$\delta'(J^{22})^{-1}\delta < 1 \tag{2.14}$$

In particular if γ is one-dimensional, narrow estimation will be superior for all estimands with $b \neq 0$ if and only if

$$|\delta| < \sqrt{J^{22}}. \tag{2.15}$$

Proof: Equation (2.13) follows directly by comparing the risk functions given by Theorems 2.4.1 and 2.4.2.

To prove (2.14), note that (2.13) clearly will be satisfied for all $b \neq 0$ if $J^{22} - \delta\delta'$ is positive definite. We shall prove that this implies (2.14). Start by the defining property of positive definiteness:

$$\forall v \neq 0 : v'(J^{22} - \delta\delta')v > 0.$$

Since the inverse square root matrix $(J^{22})^{-1/2}$ is invertible, multiplication by $(J^{22})^{-1/2}$ is a one-to-one transformation in $\mathcal{R}^q$. The above equation is thus clearly equivalent to

$$\begin{aligned}
&\forall v \neq 0 : \quad v'(J^{22})^{-1/2}(J^{22} - \delta\delta')(J^{22})^{-1/2}v > 0 \quad \Leftrightarrow \\
&\forall v \neq 0 : \quad v'v - v'(J^{22})^{-1/2}\delta\delta'(J^{22})^{-1/2}v > 0 \quad \Leftrightarrow \\
&\forall v \neq 0 : \quad \left(\delta'(J^{22})^{-1/2}v\right)^2 < v'v \quad \Leftrightarrow \\
&\forall v \neq 0 : \quad \left(\delta'(J^{22})^{-1/2}v/\|v\|\right)^2 < 1.
\end{aligned}$$

The left side of the last equation is clearly maximized by any v parallel to $(J^{22})^{-1/2}\delta$. Thus the expression above is equivalent to

$$\begin{aligned}
&\left(\delta'(J^{22})^{-1/2}(J^{22})^{-1/2}\delta/\|(J^{22})^{-1/2}\delta\|\right)^2 < 1 \quad \Leftrightarrow \\
&\delta'(J^{22})^{-1}\delta < 1.
\end{aligned}$$

□

A number of comments can be made concerning these results. The first thing to note is that for γ one-dimensional the criterion (2.15) is independent of b, and thus of the particular estimand μ under study. We could say that there is a certain "tolerance radius" around the narrow model, given by (2.15). Inside this radius, narrow estimation is better than or as good as wide estimation for all (smooth) estimands. This is the result given on p. 9 of EMMM.

In the case with γ multi- dimensional the situation is not so clear-cut. The "borderline" between narrow and wide territory will in general depend on the particular estimand under study, cf. (2.13). But there is still a smaller area given by (2.14) where narrow estimation is superior for all estimands. Correspondingly one would maybe expect that there is an area consisting of large δ-values where wide estimation is superior for all estimands. This is not so. For any given δ, one could always find an estimand with b orthogonal to δ. In such a case $b'\delta\delta'b$ is zero, and narrow estimation is more accurate.

An intuitive explanation for this phenomenon can be given. Suppose for example that γ, and thus δ, is two dimensional, and the estimand μ is increasing in both γ components. If the two δ components are now of opposite signs, the narrow model will consistently

overestimate one γ component and underestimate the other. And since μ is increasing in both components, the two errors may cancel.

We will now give a geometric interpretation of the results. Consider first the area where narrow estimation is superior for all estimands. The inequality (2.14) by definition describes the interior of a q-dimensional ellipsoid. The ellipsoid has centre in origo and "axes" that are given by the eigenvectors of $(J^{22})^{-1}$. The "half-lengths" of the axes are given by the inverse of the square-root of the corresponding eigenvalues.

The axes of the ellipsoid can be even more conveniently given in terms of J^{22}. It is a simple task to verify that the inverse of an arbitrary matrix M, has the same eigenvectors as M, and eigenvectors that are the inverses of the eigenvectors of M. Thus the axes of the ellipsoid is given by the eigenvectors of J^{22}, and has half-lengths equal to the square root of the corresponding eigenvalues of J^{22}. It can further be shown that volume of the ellipsoid is given by

$$\frac{\pi^{q/2}}{\Gamma(\frac{q}{2}+1)}|J^{22}|^{1/2}.$$

Now consider the area where narrow estimation is superior for a particular estimand. (2.13) can be written as

$$(b'\delta)^2 < b'J^{22}b.$$

The border of this set is given by

$$|b'\delta| = \sqrt{b'J^{22}b}.$$

The solution of this equation is the two hyper-planes

$$b'\delta = \pm\sqrt{b'J^{22}b}. \tag{2.16}$$

(In the case $b = 0$ the planes of course degenerate to the whole space.) Thus narrow estimation is superior for all δ between the two planes. The planes has normal vector b, and the two points on the two planes closest to origo is

$$\delta_o = \pm\frac{\sqrt{b'J^{22}b}}{b'b}b.$$

Thus the distance between the two planes is

$$2\|\delta_o\| = 2\frac{\sqrt{b'J^{22}b}}{\|b\|}.$$

We know that the ellipsoid discussed above must be contained in the area between the two planes. (Since narrow estimation is better for all estimands inside the ellipsoid it is of course also better for the particular estimand corresponding to the two planes.) We shall

demonstrate the additional fact that the two planes is tangent to the ellipsoid. Consider the two points

$$\delta_e = \pm J^{22} b (b' J^{22} b)^{-1/2}.$$

By insertion in (2.16) these points are immediately seen to be a point on each plane. Furthermore, insertion in (2.14) verifies that the two points also belong to the border of the ellipsoid. To get a characteristic picture of this situation see figure 2.16 on p. 47.

2.7 Examples; quality of the approximation

This section and the next are devoted to the study of a number of specific models. In the present section we consider some quite simple situations that will allow us to compute exact risk functions in addition to the large sample approximations. This will prove valuable in giving an idea of the n needed for the approximations to be reasonably accurate.

EXAMPLE 1 (NORMAL DENSITY)

Suppose our data has a $N(\theta, \sigma^2)$ distribution. We suspect that σ is close to some known value σ_0. The question is whether to simply use this "known" value or to estimate σ in addition to θ. To make this situation fit into our large sample framework, let $\sigma = \sigma_0\gamma$ where $\gamma = 1 + \delta/\sqrt{n}$, $(\gamma_0 = 1)$. Thus the density is given by

$$f(y|\theta,\gamma) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_0\gamma}\exp\left\{-\frac{(y-\theta)^2}{2(\sigma_0\gamma)^2}\right\}.$$

Remark: Note the parameterization $\sigma = \sigma_0\gamma$. We have chosen this parameterization to make γ, and thus δ, unscaled quantities. The theory would have worked equally well for the parameterization $\sigma = \sigma_0 + \gamma$ for example, where γ_0 in this case should be 0. Examples of different parameterizations are found in the other models studied further on.

Let us as an illustration start by verifying the regularity conditions of section 2.1. Of conditions A1-A6 the only non-trivial one is A4: By Theorem 16.8 of Billingsley (1986) an integral of the form

$$\int h(y,t)dy$$

can be differentiated under the integral sign in a neighbourhood of a point t_0 if the partial derivative of h exists and can be dominated by an integrable function $g(y)$ in the same neighbourhood. In the normal example, consider first the partial derivative with respect to γ, which is

$$\frac{\partial}{\partial\gamma}f(y|\theta,\gamma) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_0^3\gamma^4}(y-\theta)^2\exp\left\{-\frac{(y-\theta)^2}{2(\sigma_0\gamma)^2}\right\}.$$

In a neighbourhood of $\gamma_0 = 1$ the derivative is dominated by

$$\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_0^3(1/2)^4}(y-\theta)^2\exp\left\{-\frac{(y-\theta)^2}{2(\sigma_0 2)^2}\right\},$$

and this function is clearly integrable. We have thus shown that f can be differentiated under the integral sign with respect to γ. In a similar way one can show that f can be differentiated under the integral with respect to θ. The result for the second derivatives now follows by applying the same method to the first derivatives.

Now consider assumption A7: Choose the dominating function $g(y,\theta)$ to be

$$g(y,\theta) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_0 1/2}\exp\left\{-\frac{(y-\theta)^2}{2(\sigma_0 2)^2}\right\}.$$

This function goes to zero exponentially fast in y. Next, observe that $\log f(y,\theta,\gamma)$ and its derivatives will be polynomial in y. These facts readily give that all required integrals will converge.

The same reasoning can be applied to assumption A8. All logarithmic third derivatives, $\frac{\partial^3}{\partial\xi_j\partial\xi_k\partial\xi_l}\log f(y,\xi)$, will be polynomial and can be dominated by other polynomials. Thus all required integrals will converge.

Assumptions A9 and A10 are verified in much the same way: It is easy to see that both first and second derivatives of $f(y,\theta,\gamma)$ with respect to γ can be dominated by functions that go to zero at exponential rate. And since these dominating functions are only multiplied by polynomial functions, the convergence is secured. This ends the verification of the regularity conditions.

If the reader by now is bored of determining dominating functions, he may be relieved to hear that we are content to verify the regularity conditions explicitly for this one example. The conditions can be verified quite easily and with the same kind of arguments for all the other examples studied in this thesis.

Now turn to the large sample risk functions. It is a simple exercise to compute the J matrix: Take the logarithm of f, compute second order derivatives with respect to θ and γ, insert $\gamma = 1$, change sign and finally take expectation in the null distribution. The result is

$$J = \begin{pmatrix} \sigma_0^{-2} & 0 \\ 0 & 2 \end{pmatrix},$$

with corresponding inverse

$$J^{-1} = \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

From Theorem 2.6.1 we can now conclude that for all estimands, narrow estimation is

superior to wide estimation if

$$|\delta| < \sqrt{J^{22}} = \frac{1}{\sqrt{2}}.$$

Suppose now that we want to estimate the coefficient of variation,

$$\mu = \frac{\theta}{\sigma} = \frac{\theta}{\sigma_0 \gamma}.$$

In this case we obtain

$$b = J_{21} J_{22}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma} = \frac{\theta}{\sigma_0}$$

and

$$\tau_0^2 = (\frac{\partial \mu}{\partial \theta})^2 J_{11}^{-1} = 1.$$

(Remember that the partial derivatives should be evaluated at the null-point $\gamma_0 = 1$.) The risk functions are then given by:

$$\begin{aligned} r_{\text{narr}} &= \tau_0^2 + b^2\delta^2 = 1 + \frac{\theta^2}{\sigma_0^2}\delta^2, \\ r_{\text{wide}} &= \tau_0^2 + b^2 J^{22} = 1 + \frac{\theta^2}{\sigma_0^2} 1/2. \end{aligned}$$

We now compute the exact risk functions (n times mean squared error) in order to compare with the large sample approximations. Consider the wide case first. The ML-estimators for θ and σ^2 is of course the well-known $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $S^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$. So the ML-estimator for μ is

$$\hat{\mu}_{\text{wide}} = \frac{\bar{Y}}{\sqrt{S^2}}.$$

Denote the exact risk function by $r_{n,\text{wide}}(\theta, \delta)$. It is given by

$$r_{n,\text{wide}}(\theta, \delta) = En(\hat{\mu}_{\text{wide}} - \mu)^2 = E\left(\bar{Y}(S^2)^{-1/2} - \theta\sigma^{-1}\right)^2.$$

To compute this expectation, recall the following facts: $\bar{Y}$ and S^2 are independent, and $n/\sigma^2 S^2$ has a chi-square distribution with $n-1$ degrees of freedom. (Student–Fisher) Furthermore, the moments of a chi-square distributed variable can by a simple computation be shown to be

$$E(X_{n-1}^2)^\alpha = 2^\alpha \frac{\Gamma(\frac{n-1}{2} + \alpha)}{\Gamma(\frac{n-1}{2})}. \tag{2.17}$$

Using these results the answer is readily obtained:

$$r_{n,\text{wide}}(\theta,\delta) = \frac{n}{n-3} + n\left(\frac{n}{n-3} - 2\sqrt{2n}\frac{\Gamma(\frac{n-2}{2})}{\Gamma(\frac{n-1}{2})} + 1\right)\frac{\theta^2}{\sigma^2}.$$

Finally substitute $\sigma = \sigma_0(1+\delta/\sqrt{n})$.

The narrow estimator is

$$\hat{\mu}_{\text{narr}} = \bar{Y}/\sigma_0$$

and the risk function is (by simple calculations):

$$r_{n,\text{narr}}(\theta,\delta) = nE(\hat{\mu}_{\text{narr}} - \mu)^2 = (1+\delta/\sqrt{n})^2 + (1+\delta/\sqrt{n})^{-2}\frac{\theta^2}{\sigma_0^2}\delta^2.$$

Numerical comparisons of $r_{n,\text{wide}}$, r_{wide}, $r_{n,\text{narr}}$ and r_{narr} for $n = 20$, 100 and 1000 are given in figures 2.1, 2.2 and 2.3. We have chosen $\theta = 10$ and $\sigma_0 = 2$ in the plots, but other choices give comparable results. The numerical values plotted are the square roots of the risk functions, cf. the comment on page 9. □

EXAMPLE 2 (TWO EXPONENTIAL VARIABLES)

Now consider a vector-valued variable $Y = (V, W)'$. Let V and W be independent and exponentially distributed with parameters θ and λ. As a motivation consider a technological system consisting of two independent components with exponential failure times. Suppose one wants to estimate various parameters of the system based on collected failure data $(V_i, W_i)'$. Suppose further that the two components are quite similar so that it is reasonable to expect that λ is close to θ. The question is whether to postulate $\lambda = \theta$, or to estimate both parameters. We shall investigate this question in our large sample framework and start by letting $\lambda = \theta\gamma$, where $\gamma = 1 + \delta/\sqrt{n}$. This gives the density

$$f(v,w|\theta,\gamma) = \theta\exp\{-\theta v\}\theta\gamma\exp\{\theta\gamma w\}.$$

The J matrix is then computed to be:

$$J = \begin{pmatrix} 2/\theta^2 & 1/\theta \\ 1/\theta & 1 \end{pmatrix}.$$

The lower right corner of the inverse is $J^{22} = 2$, from which we immediately conclude that narrow estimation (postulating $\lambda = \theta$) is better than wide estimation for all estimands with $b \neq 0$ if

$$|\delta| < \sqrt{2}.$$

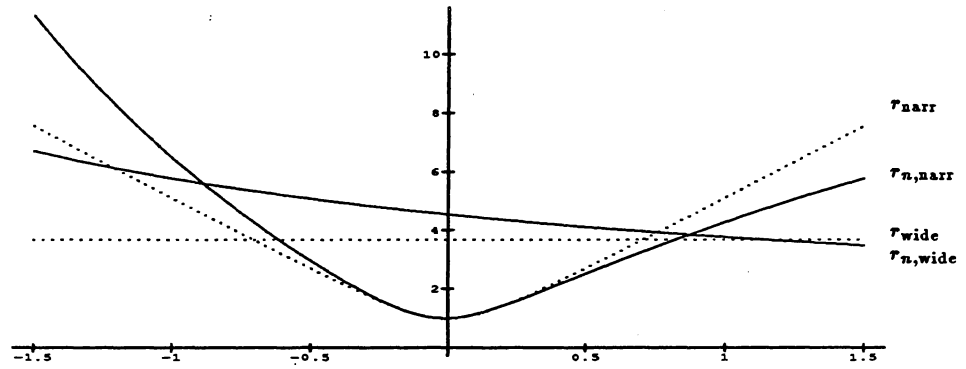

Figure 2.1: Square root of risk functions plotted as a function of δ: Normal density, estimand $\mu = \theta/\sigma$. Large sample approximations shown with dotted lines, and exact values for $n = 20$ shown with solid lines.

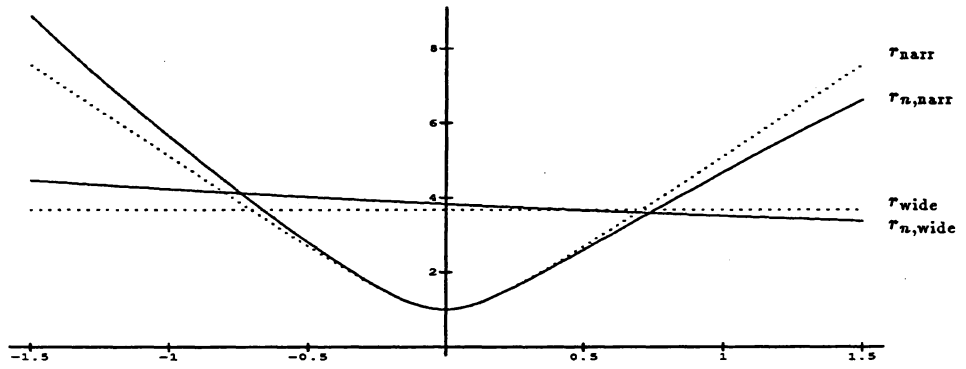

Figure 2.2: Square root of risk functions: Normal density, estimand $\mu = \theta/\sigma$, $n = 100$.

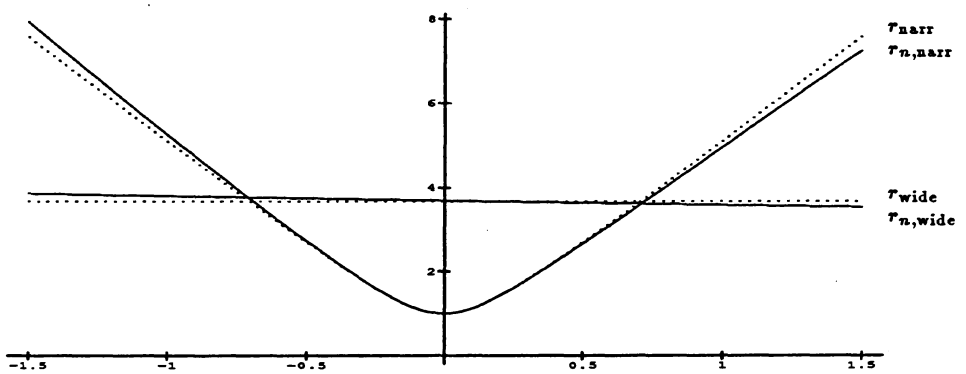

Figure 2.3: Square root of risk functions: Normal density, estimand $\mu = \theta/\sigma$. $n = 1000$.

Now let us consider some specific estimands. Suppose for example that the system under study fails as soon as one of the components fails, and the desired estimand is expected time to failure which is:

$$\mu = \frac{1}{\theta + \lambda}.$$

(Recall that the minimum of two independent exponentials is exponential with the sum of the two parameters as parameter.) It is now a trivial exercise to compute b, and the answer is 0. Consequently narrow and wide estimation are large sample equivalent in this situation. We conclude that it may very well happen that $b = 0$ in natural examples even though μ depends on γ. (The most trivial example on $b = 0$ is an estimand μ independent of γ and $J_{21} = 0$.) We invite the reader to check that $b = 0$ also happens for the following estimands:

$$E \max(V, W) = \frac{1}{\theta} + \frac{1}{\lambda} - \frac{1}{\theta + \lambda},$$

$$E(V + W) = \frac{1}{\theta} + \frac{1}{\lambda}$$

and

$$EVW = \frac{1}{\theta\lambda}.$$

Let us instead consider the estimation of the expectation of W alone:

$$\mu = \frac{1}{\lambda} = \frac{1}{\theta\gamma}.$$

In this case we obtain

$$\tau_0^2 = \frac{1}{2\theta^2}, \quad b = \frac{1}{2\theta},$$

which gives the risk functions:

$$\begin{aligned} r_{\text{wide}}(\theta, \delta) &= \frac{1}{\theta^2}, \\ r_{\text{narr}}(\theta, \delta) &= \frac{1}{2\theta^2} + \frac{1}{4\theta^2}\delta^2. \end{aligned}$$

Let us now compare this to the exact risk functions. The wide estimator for λ is $\hat{\lambda}_{\text{wide}} = n/\sum_{i=1}^{n} W_i$, which gives the corresponding estimator for μ:

$$\hat{\mu}_{\text{wide}} = 1/\hat{\lambda}_{\text{wide}} = \frac{1}{n}\sum_{i=1}^{n} W_i. \tag{2.18}$$

(Remember the invariance property of ML-estimators. We may just as well determine the ML-estimator of μ in the parameterization (θ, λ) as in the parameterization (θ, γ).) The risk function is now easily determined remembering that the sum of exponential variables is gamma distributed:

$$r_{n,\text{wide}}(\theta, \delta) = \frac{1}{\theta^2(1 + \delta/\sqrt{n})^2}.$$

The narrow estimator for θ is clearly $\hat{\theta}_{\text{narr}} = 2n/(\sum_{i=1}^n V_i + \sum_{i=1}^n W_i)$, with corresponding μ estimator

$$\hat{\mu}_{\text{narr}} = \frac{1}{2n}(\sum_{i=1}^n V_i + \sum_{i=1}^n W_i). \tag{2.19}$$

This gives the risk function (using again the gamma distribution of the estimator):

$$r_{n,\text{narr}}(\theta, \delta) = \frac{\delta^2 + 2\delta\sqrt{n} + 2n + \delta^2 n}{4\delta^2\theta^2 + 8\delta\sqrt{n}\theta^2 + 4n\theta^2}.$$

Numerical comparisons of exact and large sample risk functions for $\theta = 0.1$ and $n = 20$, 100 and 1000 are given in figures 2.4, 2.5 and 2.6. □

EXAMPLE 3 (MILD REGRESSION)

Suppose that we have observations Y_i that we hope are i.i.d. What will happen if we perform the analysis based on the i.i.d. assumption, and the Y_i's really have a mild regression structure? Suppose the distribution of Y_i depends on some covariate x_i. We model this by saying that

$$Y_i = \eta + \gamma x_i + \sigma\varepsilon_i$$

where the ε_i is independent $N(0,1)$-variables. Let $\gamma_0 = 0$, so that $\gamma = \delta/\sqrt{n}$. Corresponding to the view taken in the general theory, we shall model the x_i as random, and suppose that they are independent of each other and of the ε_i's. We shall further make the assumption that the x_i's have a $N(0, \tau^2)$ distribution. If the x_i's are unknown, the situation is really not changed, since they may be absorbed in the remainder terms and have no effect except to increase the variance. But if the x_i's are known, we could utilize them in the estimation. Let us now use our general theory to evaluate the gain or loss obtained by including the x_i's in the estimation procedure: The conditional density of the Y_i's is: (let $\theta = (\eta, \sigma)'$)

$$f(y|\theta, \gamma, x) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left\{-\frac{(y - \eta - \gamma x)^2}{2\sigma^2}\right\}.$$

The corresponding conditional J matrix is determined as before by computing all second derivatives of the logarithm of f, substituting $\gamma = 0$, changing sign and taking expectation.

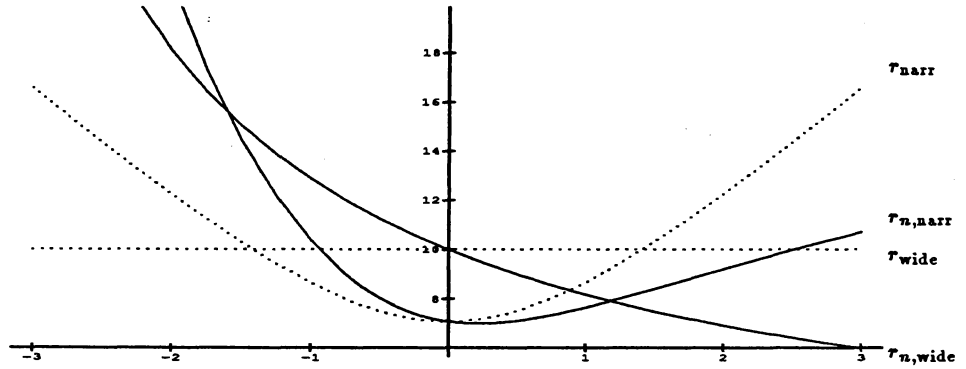

Figure 2.4: Square root of risk functions: Two exponential variables, estimand $\mu = 1/\lambda$, $n = 20$.

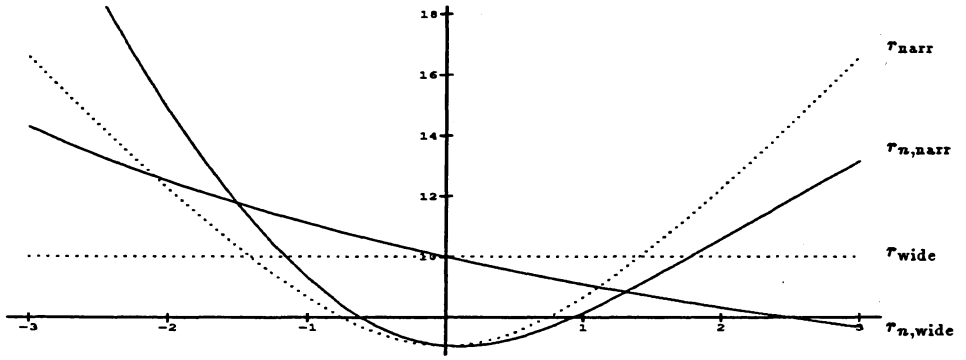

Figure 2.5: Square root of risk functions: Two exponential variables, estimand $\mu = 1/\lambda$, $n = 100$.

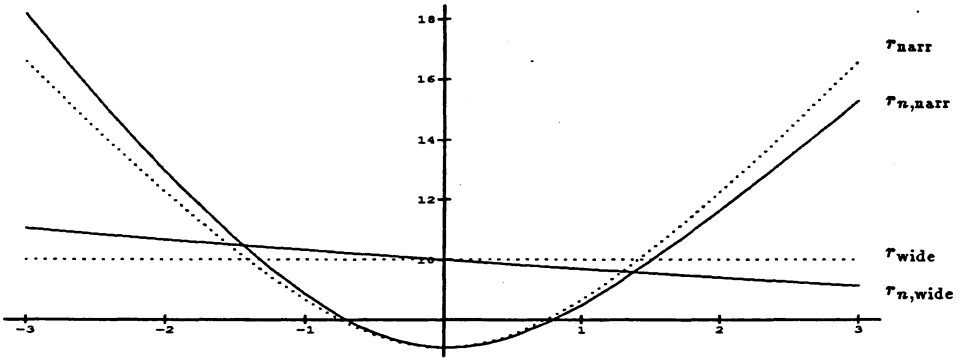

Figure 2.6: Square root of risk functions. Two exponential variables, estimand $\mu = 1/\lambda$, $n = 1000$.

The expectation should now be in the conditional distribution of Y given x and with the parameter $\gamma = 0$. These are simple computations and the result is:

$$J(x) = \begin{pmatrix} \sigma^{-2} & 0 & x\sigma^{-2} \\ 0 & 2\sigma^{-2} & 0 \\ x\sigma^{-2} & 0 & x^2\sigma^{-2} \end{pmatrix}.$$

Taking expectation in the distribution of x gives the unconditional J-matrix:

$$J = \begin{pmatrix} \sigma^{-2} & 0 & 0 \\ 0 & 2\sigma^{-2} & 0 \\ 0 & 0 & \tau^2\sigma^{-2} \end{pmatrix}.$$

This gives $J^{22} = \tau^{-2}\sigma^2$, and we conclude that narrow estimation is better for all estimands with $b \neq 0$ if

$$|\delta| < \frac{\sigma}{\tau}.$$

Suppose now that the estimand under study is a linear function of η and γ

$$\mu = a_1\eta + a_2\gamma.$$

The aim could for example be to predict a future observable for given x. (Set $a_1 = 1$ and $a_2 = x$.) The general theory then easily gives $b = -a_2$ and $\tau_0^2 = a_1^2\sigma^2$, which in turn gives:

$$\begin{aligned} r_{\text{wide}}(\theta, \delta) &= a_2^2\frac{\sigma^2}{\tau^2} + a_1^2\sigma^2. \\ r_{\text{narr}}(\theta, \delta) &= a_2^2\delta^2 + a_1^2\sigma^2. \end{aligned}$$

Once more we want to compare these large sample approximations to the exact values. Consider the wide estimator first. Define

$${}_nY = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad {}_nx = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad {}_n\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Define further

$$C = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \eta \\ \gamma \end{pmatrix}, \quad a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}.$$

The model can now be written on matrix form:

$$ {}_nY = C\beta + \sigma\, {}_n\varepsilon. $$

From the theory of linear normal models the least squares and ML-estimator of β is given by

$$ \hat{\beta} = (C'C)^{-1}C'\, {}_nY. $$

The estimator is unbiased and has variance $(C'C)^{-1}\sigma^2$. In our setup these are the conditional moments given ${}_nx$.

We are now ready to determine the exact risk function:

$$ \begin{aligned} r_{n,\text{wide}}(\theta,\delta) &= nE(\hat{\mu}_{\text{wide}} - \mu)^2 = nE(a'\hat{\beta} - a'\beta)^2 = \\ & na'EE\left((\hat{\beta}-\beta)(\hat{\beta}-\beta)'\Big|\, {}_nx\right) a = na'E(C'C)^{-1}a\sigma^2. \end{aligned} $$

The matrix $(C'C)^{-1}$ is easily shown to be

$$ \frac{1}{n}\frac{1}{S_x^2}\begin{pmatrix} S_x^2+\bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}, $$

where $\bar{x}$ and $S_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$ are empirical mean and variance of the x_i's. Using Student–Fisher's result and (2.17) on p. 31 the expectation of $(C'C)^{-1}$ is now easily determined. The final answer is

$$ r_{n,\text{wide}} = a_2^2 \frac{n}{n-3}\frac{\sigma^2}{\tau^2} + a_1^2 \frac{n-2}{n-3}\sigma^2. $$

The narrow estimator of θ is of course $\hat{\theta}_{\text{narr}} = \bar{Y}$ with corresponding μ estimator $\hat{\mu} = a_1\bar{Y}$. The exact risk function of this estimator is found by a simple calculation using the rule of double expectation. The result is:

$$ r_{n,\text{narr}} = a_2^2\delta^2 + a_1^2\left(\sigma^2 + \frac{\delta^2}{n}\tau^2\right). $$

Numerical plots of exact and large sample risk functions for $a_1 = 1$, $a_2 = 2$, $\sigma = 3$, $\tau = 4$ and $n = 10$, 20 and 100 are given in figures 2.7, 2.8 and 2.9. □

EXAMPLE 4 (LOGNORMAL DENSITY)

Let us now try a distribution with heavy tails: Suppose that the data have a lognormal

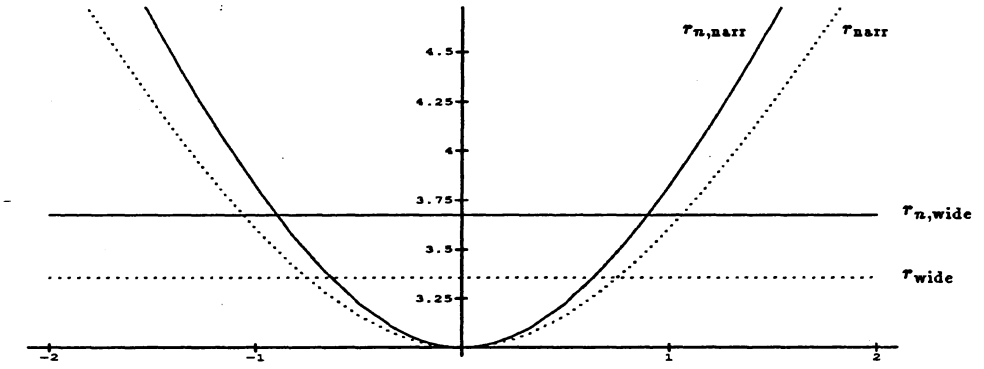

Figure 2.7: Square root of risk functions: Mild regression, estimand $\mu = \theta + 2\gamma$, $n = 10$.

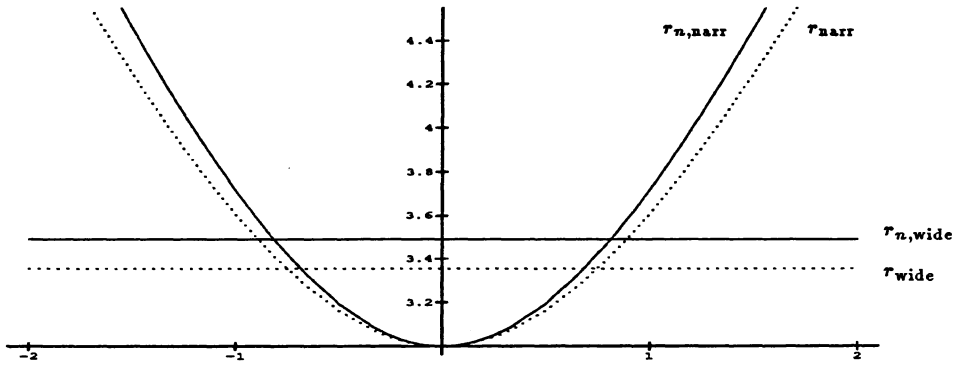

Figure 2.8: Square root of risk functions: Mild regression, estimand $\mu = \theta + 2\gamma$, $n = 20$.

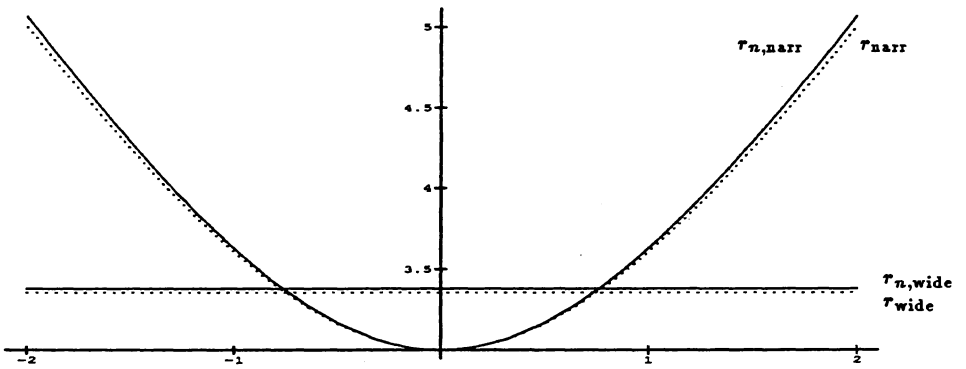

Figure 2.9: Square root of risk functions: Mild regression, estimand $\mu = \theta + 2\gamma$, $n = 100$.

distribution with parameters θ and σ, that is with density

$$\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\frac{1}{y}f(y|\theta,\sigma) = \exp\left\{-\frac{(\log y - \theta)^2}{2\sigma}\right\}.$$

We want to study this model when σ is close to some known value σ_0 and model this by letting $\sigma = \sigma_0\gamma$ where $\gamma = 1+\delta/\sqrt{n}$. By the defining property of the lognormal distribution, the logarithm of a lognormally distributed variable will have a normal distribution with the same parameters. Using this it is a simple task to compute the J matrix, which turns out to be the same as in the normal example:

$$J = \begin{pmatrix} \sigma_0^{-2} & 0 \\ 0 & 2 \end{pmatrix}.$$

Suppose now that our estimand is the expectation in the lognormal distribution

$$\mu = \exp\left\{\theta + \frac{1}{2}\sigma^2\right\}.$$

By simple computations the large sample risk functions are found to be

$$\begin{aligned} r_{\text{wide}}(\theta,\delta) &= \exp\{2\theta+\sigma_0^2\}\sigma_0^2\left(1+\frac{1}{2}\sigma_0^2\right), \\ r_{\text{narr}}(\theta,\delta) &= \exp\{2\theta+\sigma_0^2\}\sigma_0^2(1+\delta^2\sigma_0^2). \end{aligned}$$

We now turn to the exact risk functions. Define $Z_i = \log Y_i$. By writing out the likelihood in terms of Z_i, it is easily seen that the maximizing estimators will have the same form as the estimators in the normal situation: $\hat{\theta}_{\text{wide}} = \bar{Z}$, $\hat{\sigma}^2_{\text{wide}} = \frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})^2$ and $\hat{\theta}_{\text{narr}} = \bar{Z}$. Since Z_i is normally distributed, the estimators will of course also have the same distribution as in the normal case. The computation of exact risk function for the corresponding μ-estimators

$$\hat{\mu}_{\text{wide}} = \exp\left\{\hat{\theta}_{\text{wide}} + \frac{1}{2}\hat{\sigma}^2_{\text{wide}}\right\}$$

and

$$\hat{\mu}_{\text{narr}} = \exp\left\{\hat{\theta}_{\text{narr}} + \frac{1}{2}\sigma_0^2\right\}$$

can now be accomplished. Student–Fisher's theorem and the expressions for moment-generating functions in the normal and chi-square distributions will do the job. The answer

is

$$r_{n,\text{wide}}(\theta,\delta) = ne^{2\theta}\left[e^{\sigma^2} + \frac{\exp\left\{\frac{2}{n}\sigma^2\right\}}{\left(1-\frac{2}{n}\sigma^2\right)^{\frac{n-1}{2}}} - 2\frac{\exp\left\{\frac{n+1}{2n}\sigma^2\right\}}{\left(1-\frac{1}{n}\sigma^2\right)^{\frac{n-1}{2}}}\right]$$

and

$$r_{n,\text{narr}}(\theta,\delta) = ne^{2\theta}\left[e^{\sigma^2} + \exp\left\{\sigma_0^2 + \frac{2}{n}\sigma^2\right\} - 2\exp\left\{\frac{1}{2}\sigma_0^2 + \frac{n+1}{2n}\sigma^2\right\}\right].$$

(Remember that $\sigma = \sigma_0(1+\delta/\sqrt{n})$.)

Numerical plots of exact and large sample risk functions for $\sigma_0 = 2$, $\theta = 3$ and $n = 20$, 100 and 1000 are given in figures 2.10, 2.11 and 2.12. Note that the convergence seems to be rather slow in this example. One could of course argue that this is exactly as expected due to the heavy tails of the lognormal distribution. But the estimators of θ and σ after all turned out to have exactly the same distribution as in the normal example. We would therefore prefer to draw attention to the extreme unlinearity of the estimand $\mu = \exp\{\theta + \frac{1}{2}\sigma^2\}$. □

We have now considered four different examples and compared large sample approximations and exact risk functions. Can we make any general statements?

It seems obvious that no exact statement about the rapidity of convergence can be made just from the study of a few examples. Further, the rapidity of convergence is clearly seen to vary between the examples studied, and in particular between different estimands. (It should not be difficult to construct artificial examples showing an arbitrary slow rate of convergence.) The rapidity of convergence also varies with the magnitude of δ, not unexpectedly being slower for large δ than for small. (In most cases, one could for given n, always find a sufficiently large δ to make the approximation arbitrarily bad. This is not a very serious problem, however, since one would usually only be interested in evaluating the risk function out to a certain point beyond the tolerance radius.)

In spite of these objections, and without giving a more precise content to the notion "rapidity of convergence", we shall try to give a few rules of thumb.

- The approximation may in some situations be useful for as small n as 20. Note that the tolerance radius of the narrow model is quite exactly approximated for $n = 20$ in all examples except the lognormal one.
- The approximation should be good for most purposes, for n as large as 100.
- The approximation will usually be excellent for n as large as 1000.
- The approximation can be expected to be better for approximately linear μ functions (cf. the mild regression example) than for highly unlinear ones (cf. the lognormal example). In particular the approximation can not be expected to be good if $\delta/\sqrt{n}$ is outside an approximately linear neighbourhood of μ around γ^0.

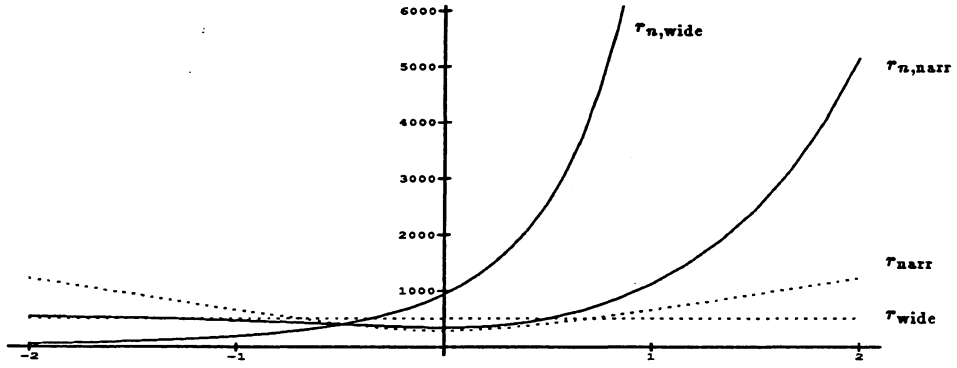

Figure 2.10: Square root of risk functions: Lognormal density, $n = 20$, estimand $\mu = \exp\left\{\theta + \frac{1}{2}\sigma^2\right\}$.

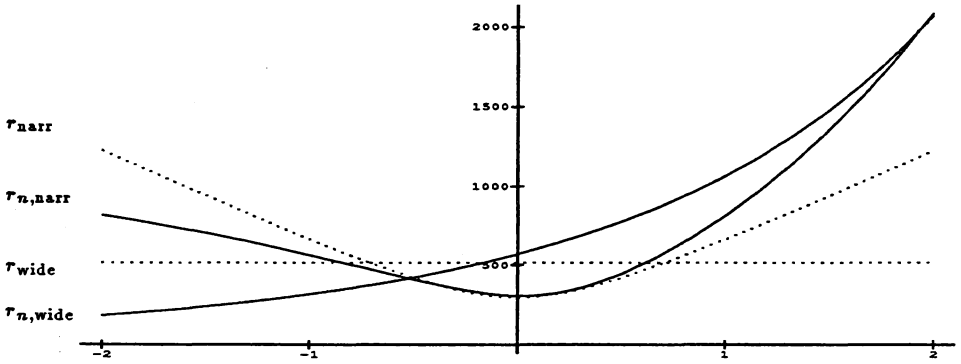

Figure 2.11: Square root of risk functions: Lognormal density, $n = 100$, estimand $\mu = \exp\left\{\theta + \frac{1}{2}\sigma^2\right\}$.

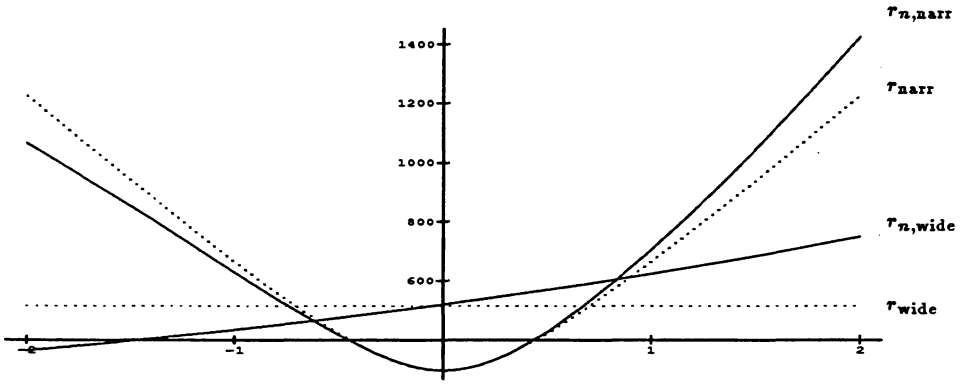

Figure 2.12: Square root of risk functions: Lognormal density, $n = 1000$, estimand $\mu = \exp\left\{\theta + \frac{1}{2}\sigma^2\right\}$.

Before leaving this topic, let us take time for one general remark. In my view the most important property of large sample approximations is probably to simplify matters sufficiently to give a qualitative understanding of a complex phenomenon. When the goal is to obtain a concrete number for some specific use, I believe that it will usually be simpler to use an exact numerical method, like numerical integration or stochastic simulation, than to check the exactness of a large sample approximation. This remark applies to all large sample approximations, of course, not only the one presented in this thesis.

2.8 Examples; multi-dimensional deviations

We now turn to the study of some examples with possible model departures in several directions. That is with multi-dimensional δ.

EXAMPLE 5 (REGRESSION WITH QUADRATICITY AND VARIANCE HETEROGENEITY)

Consider a situation where the null model is a standard linear regression model, with data Y_i and covariates x_i. Suppose that the following violations of the model are suspected:

- The expectation Y_i may not be linear in x_i.
- The variance of Y_i may depend on x_i.

As one model for this situation we choose a regression model with a quadratic term and a variance that is multiplied by a factor depending on x_i. Specifically we assume the density

$$f(y|\theta,\gamma,x) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma(1+\nu x)}\exp\left\{-\frac{(y-\alpha-\beta x-\varepsilon\sigma x^2)^2}{2\sigma^2(1+\nu x)^2}\right\}.$$

Here $\theta = (\alpha,\beta,\sigma)'$ and $\gamma = (\varepsilon,\nu)'$. The null model corresponds to $\gamma^0 = (0,0)'$. As in the mild regression example, we shall assume that the x_i's has a $N(0,\tau^2)$ distribution. Computation of the J matrix is not more complicated than before, it only has more elements! Computing second derivatives and taking expectation in the conditional null-distribution of Y given x gives:

$$J(x) = \begin{pmatrix} \frac{1}{\sigma^2} & \frac{x}{\sigma^2} & 0 & \frac{x^2}{\sigma} & 0 \\ \frac{x}{\sigma^2} & \frac{x^2}{\sigma^2} & 0 & \frac{x^3}{\sigma} & 0 \\ 0 & 0 & \frac{2}{\sigma^2} & 0 & 0 \\ \frac{x^2}{\sigma} & \frac{x^3}{\sigma} & 0 & x^4 & 0 \\ 0 & 0 & \frac{2x}{\sigma} & 0 & 2x^2 \end{pmatrix}.$$

The unconditional J is the expectation of this matrix, in distribution of x. From results

about the normal distribution we know that $Ex^3 = 0$ and $Ex^4 = 3\tau^4$. This gives the result

$$J = \begin{pmatrix} \frac{1}{\sigma^2} & 0 & 0 & \frac{\tau^2}{\sigma} & 0 \\ 0 & \frac{\tau^2}{\sigma^2} & 0 & 0 & 0 \\ 0 & 0 & \frac{2}{\sigma^2} & 0 & 0 \\ \frac{\tau^2}{\sigma} & 0 & 0 & 3\tau^4 & 0 \\ 0 & 0 & 0 & 0 & 2\tau^2 \end{pmatrix}.$$

The inverse is

$$J^{-1} = \begin{pmatrix} \frac{3\sigma^2}{2} & 0 & 0 & -\frac{\sigma}{2\tau^2} & 0 \\ 0 & \frac{\sigma^2}{\tau^2} & 0 & 0 & 0 \\ 0 & 0 & \frac{\sigma^2}{2} & 0 & 0 \\ -\frac{\sigma}{2\tau^2} & 0 & 0 & \frac{1}{2\tau^4} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2\tau^2} \end{pmatrix}.$$

The ellipse defining narrow superiority for all estimands (given by (2.14) on p. 26), can now be easily determined:

$$\delta'(J^{22})^{-1}\delta = \delta' \begin{pmatrix} 2\tau^4 & 0 \\ 0 & 2\tau^2 \end{pmatrix} \delta = 2\tau^4\delta_1^2 + 2\tau^2\delta_2^2.$$

The ellipse is thus defined by

$$2\tau^4\delta_1^2 + 2\tau^2\delta_2^2 = 1.$$

This is the equation of an ellipse with axes parallel to the coordinate axes. The half length in δ_1 direction (corresponding to quadraticity), is $\frac{1}{\sqrt{2}\tau^2}$. And the half length in δ_2 direction (corresponding to variance heterogeneity), is $\frac{1}{\sqrt{2}\tau}$.

Let us now consider some specific estimands. A natural estimand to consider is a linear function of α, β and $\varepsilon\sigma$:

$$\mu = a_1\alpha + a_2\beta + a_3\varepsilon\sigma.$$

For example, to predict a future observable for given x, it would be natural to use an estimator of the form above with $a_1 = 1$, $a_2 = x$ and $a_3 = x^2$.

For this estimand we obtain

$$b = \begin{pmatrix} a_1\tau^2 - a_3 \\ 0 \end{pmatrix} \sigma, \quad \text{and} \quad \tau_0^2 = a_1^2\sigma^2 + a_2^2\frac{\sigma^2}{\tau^2},$$

which gives the risk functions

$$r_{\text{wide}}(\theta, \delta) = (a_1\tau^2 - a_3)^2\frac{\sigma^2}{2\tau^4} + a_1^2\sigma^2 + a_2^2\frac{\sigma^2}{\tau^2},$$

$$r_{\text{narr}}(\theta,\delta) = (a_1\tau^2 - a_3)^2\sigma^2\delta_1^2 + a_1^2\sigma^2 + a_2^2\frac{\sigma^2}{\tau^2}.$$

Note that the narrow risk function is independent of δ_2, which is a consequence of the fact that the second component of b is 0. We thus conclude that for the estimand in question we could safely use the narrow estimator although moderate variance heterogeneity is suspected. On the other hand, moderate quadraticity is a more serious departure.

The two lines defining the boundary of narrow supremacy territory are by (2.16) given by

$$b'\delta = \pm\sqrt{b'J^{22}b}.$$

Substituting the above expressions for b and τ_0^2 and simplifying gives

$$\delta_1 = \pm\frac{1}{\sqrt{2}\tau^2}.$$

Numerical plots giving characteristic visualizations of these results are given in figures 2.13 and 2.14.

Let us now consider another estimand. We choose the coefficient of variation for a given x value a. That is

$$\mu = \frac{\alpha + a\beta + a^2\varepsilon\sigma}{(1+a\nu)\sigma}.$$

In this case we obtain

$$b = \begin{pmatrix} \tau^2 - a^2 \\ \frac{a(\alpha+a\beta)}{\sigma} \end{pmatrix}, \quad \text{and} \quad \tau_0^2 = 1 + \frac{(\alpha+a\beta)^2}{2\sigma^2} + \frac{a^2}{\tau^2},$$

with corresponding risk functions

$$r_{\text{wide}}(\theta,\delta) = \frac{(\tau^2-a^2)^2}{2\tau^4} + \frac{a^2(\alpha+a\beta)^2}{2\sigma^2\tau^2} + 1 + \frac{(\alpha+a\beta)^2}{2\sigma^2} + \frac{a^2}{\tau^2},$$

$$r_{\text{narr}}(\theta,\delta) = \frac{\left((a^2-\tau^2)\sigma\delta_1 - a(\alpha+a\beta)\delta_2\right)^2}{\sigma^2} + 1 + \frac{(\alpha+a\beta)^2}{2\sigma^2} + \frac{a^2}{\tau^2}.$$

Numerical plots of these results are given in figures 2.15 and 2.16. □

EXAMPLE 6 (GOMPERTZ–MAKEHAM HAZARD RATE)

In life insurance, a common model used to describe mortality is a Gompertz–Makeham hazard rate. It has the form

$$h(y) = \alpha + \beta c^y.$$

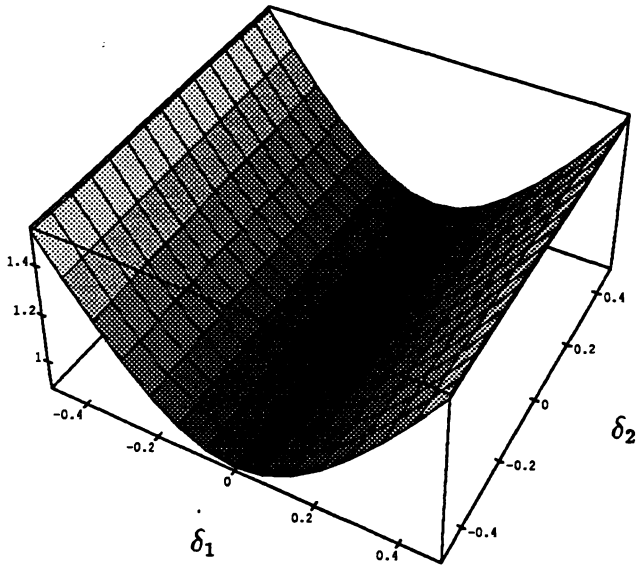

Figure 2.13: Regression with quadraticity and variance heterogeneity. The relative risk of the narrow estimator $\sqrt{r_{\text{narr}}}/\sqrt{r_{\text{wide}}}$ is plotted as a function of δ_1 and δ_2. The estimand is $\mu = a_1\alpha + a_2\beta + a_3\varepsilon\sigma$. Numerical values used in the plot are $a_1 = 1$, $a_2 = 3$, $a_3 = 9$, $\sigma = 1$ and $\tau = 2$.

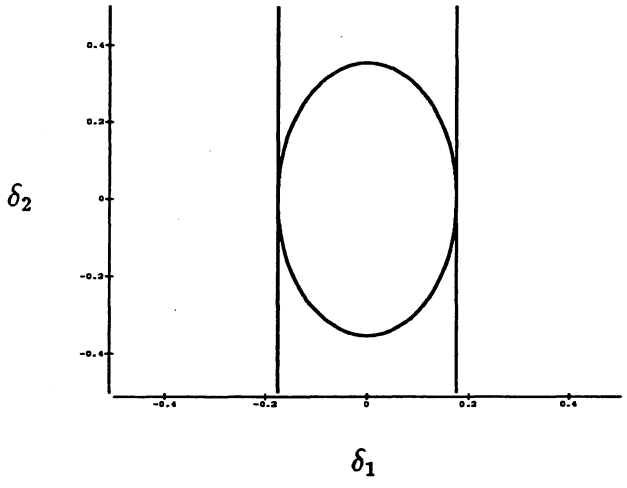

Figure 2.14: Regression with quadraticity and variance heterogeneity. The ellipse shows the area where narrow estimation is better than wide estimation for all estimands. The space between the two lines is the area where the narrow estimator is better than the wide for the estimand $\mu = a_1\alpha + a_2\beta + a_3\varepsilon\sigma$. Numerical values as in the plot above.

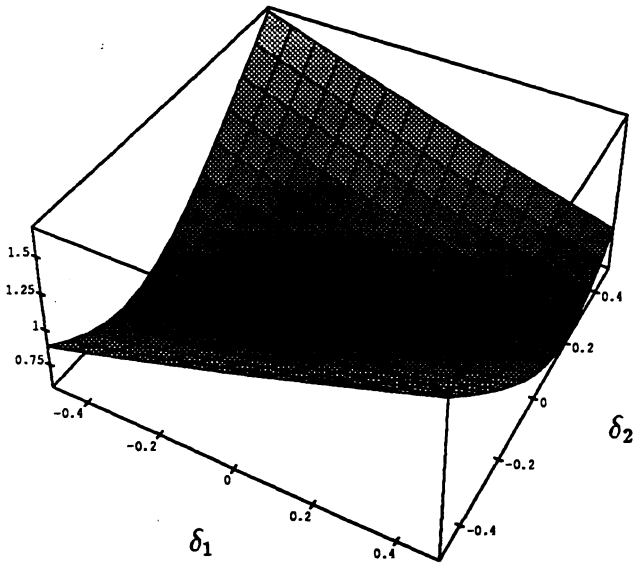

Figure 2.15: Regression with quadraticity and variance heterogeneity. Relative risk of the narrow estimator $\sqrt{r_{\text{narr}}}/\sqrt{r_{\text{wide}}}$ plotted as a function of δ_1 and δ_2. Estimand $\mu = \big(\alpha + a\beta + a^2\varepsilon\sigma\big)/\big((1 + a\nu)\sigma\big)$. Numerical values used in the plot are $a = 3$, $\alpha = 1$, $\beta = 1$, $\sigma = 1$ and $\tau = 2$.

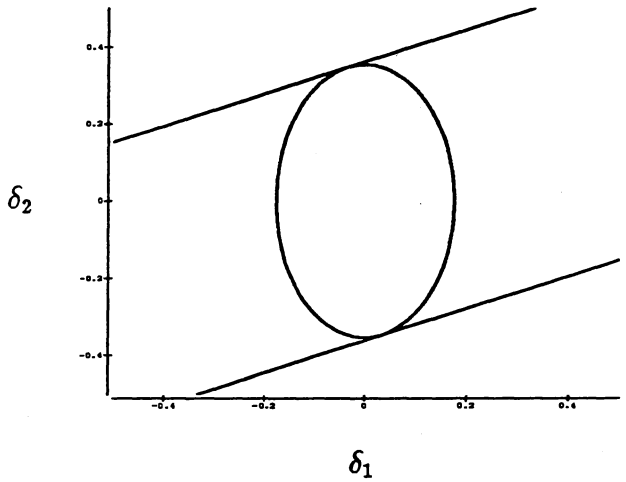

Figure 2.16: Regression with quadraticity and variance heterogeneity. The ellipse shows the area where narrow estimation is better than wide estimation for all estimands. The space between the two lines is the area where the narrow estimator is better than the wide for the estimand $\mu = \big(\alpha + a\beta + a^2\varepsilon\sigma\big)/\big((1 + a\nu)\sigma\big)$. Numerical values as in the plot above.

We find it more convenient to reparameterize it as

$$h(y) = \alpha + \beta \exp(ky),$$

where $k = \log c$.

Suppose now that one wanted to estimate the mortality in a small sub-population. In collective life insurance for example, it might be of interest to estimate the mortality in a certain industry. If the data are quite scarce, the estimation of the three parameters in the hazard rate might be too inaccurate. A narrow model that could be proposed is to take the nation-wide estimated values of β and k as given, and only estimate α. This model would fix the shape of the mortality curve, but allow the level to vary.

Suppose, for example, that the population under study has higher mortality than normal as a result of industry specific dangers. It may then be reasonable to believe that the increase in the mortality would not depend of age, thus agreeing with the proposed model. The nation-wide mortality estimates would of course be quite accurate due to the large amount of data available.

We shall as usual compare these two competing models in our large sample framework. Start by setting $\beta = \gamma_1\beta_0$ and $k = \gamma_2 k_0$, where β_0 and k_0 are the given values. In our general notation θ corresponds to α and $\gamma = (\gamma_1, \gamma_2)'$. The narrow model corresponds to $\gamma^0 = (1,1)'$.

To determine the J matrix, we first need to find the density corresponding to this hazard rate. The cumulative distribution function is given by

$$\begin{aligned} F(y|\theta,\gamma) &= 1 - \exp\left\{-\int_0^y h(t)\,dt\right\} \\ &= 1 - \exp\left\{\frac{\gamma_1\beta_0}{\gamma_2 k_0}(1 - e^{\gamma_2 k_0 y}) - \alpha y\right\}. \end{aligned}$$

Derivation thus gives the density

$$f(y|\theta,\gamma) = \left(\alpha + \gamma_1\beta_0 e^{\gamma_2 k_0 y}\right) \exp\left\{\frac{\gamma_1\beta_0}{\gamma_2 k_0}(1 - e^{\gamma_2 k_0 y}) - \alpha y\right\}.$$

Taking the logarithm and computing second derivatives of this density gives quite messy expressions, and the ensuing expectation can not be determined symbolically. However, there is no problem computing the expectation by numerical integration.

Let us now decide to use the values $\alpha = 0.0009$, $\beta_0 = 0.000044$ and $k_0 = \log 10^{0.042}$. These values are taken from a mortality table used in practice in Norwegian collective life insurance. In this case numerical integration gives

$$J = \begin{pmatrix} 34076.9 & 40.5117 & 199.027 \\ 40.5117 & 0.899469 & 6.47134 \\ 199.027 & 6.47134 & 47.7911 \end{pmatrix}.$$

The inverse is

$$J^{-1} = \begin{pmatrix} 0.0000394969 & -0.023094 & 0.00296264 \\ -0.023094 & 56.6176 & -7.57034 \\ 0.00296264 & -7.57034 & 1.03368 \end{pmatrix}.$$

Let us use this to determine the ellipse where narrow estimation is better than wide estimation for all estimands. The eigenvectors of J^{22} are

$$v_1 = \begin{pmatrix} 0.991 \\ -0.133 \end{pmatrix} \quad \text{and} \quad v_2 = \begin{pmatrix} 0.133 \\ 0.991 \end{pmatrix},$$

giving the direction of the axes of the ellipse. The square root of the corresponding eigenvalues are 7.59 and 0.145, giving the half lengths of the axes.

Note that the ellipse is very much longer in one direction than in the other. In trying to give an intuitive explanation of this phenomenon, we plot the density for different parameter values. See the figure 2.17.

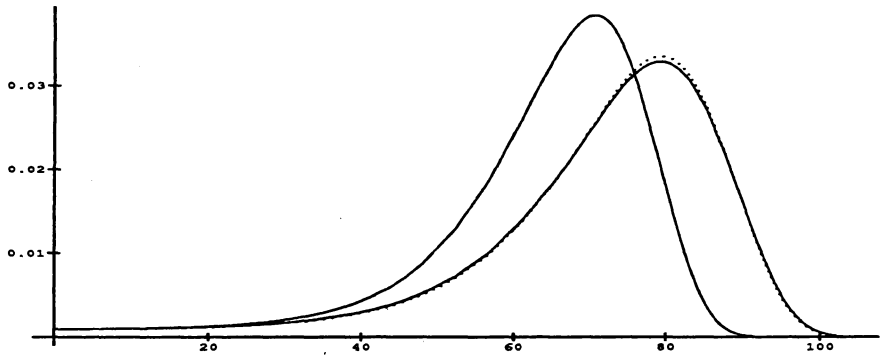

Figure 2.17: Densities from the Gompertz–Makeham example. The null density corresponding to $\gamma = (0, 0)'$ shown with a dotted line. The solid line almost covering the dotted one corresponds to $\gamma = v_1/50$ and the last solid line corresponds to $\gamma = v_2/50$. (The eigenvectors v_1 and v_2 have unit length.) Other numerical values as in the main text.

Note from the plot that a deviation a fixed distance from the null point along v_1 gives an almost unchanged density. For comparison a deviation the same distance along v_2 gives a quite different density. We would thus expect the wide estimator of γ to have a large variance in the direction of v_1.

Now consider two possibilities: If μ varies little when γ varies in v_1 direction, it is clear that the narrow model can tolerate a large deviation in this direction. On the other hand, if μ varies much when γ varies in v_1 direction, a deviation in this direction will give a large risk for the narrow estimator. But in this case the wide estimator will have a large risk too, and the narrow estimator might be better after all.

In contrast, a deviation in v_2 direction will be precisely detected by the wide estimator,

making the narrow estimator a poor alternative for μ functions sensitive to deviations in this direction.

Before leaving this topic, we question the suitability of the Gompertz–Makeham model itself on the grounds that widely spaced parameter points have almost equal densities. (This does not only apply to the densities, plots of the hazard rates show corresponding results.) Maybe some two parametric model could give almost the same spectre of densities?

We now turn to consider a specific estimand. We choose to estimate $h(40)$, the mortality for a 40-year-old person. (This estimand would be a close approximation to the net unit premium for a one year life insurance of a 40-year-old person.)

This estimand has $b = (-9.171 \cdot 10^{-4}, -2.306 \cdot 10^{-3})'$ and $\tau_0^2 = 2.935 \cdot 10^{-5}$, with corresponding risk functions:

$$r_{\text{wide}} = 5.044 \cdot 10^{-5}.$$

$$r_{\text{narr}} = 2.935 \cdot 10^{-5} + 8.411 \cdot 10^{-7}\delta_1^2 + 4.230 \cdot 10^{-6}\delta_1\delta_2 + 5.318 \cdot 10^{-6} \cdot \delta_2^2.$$

Plots of these results are given in figures 2.18 and 2.19. □

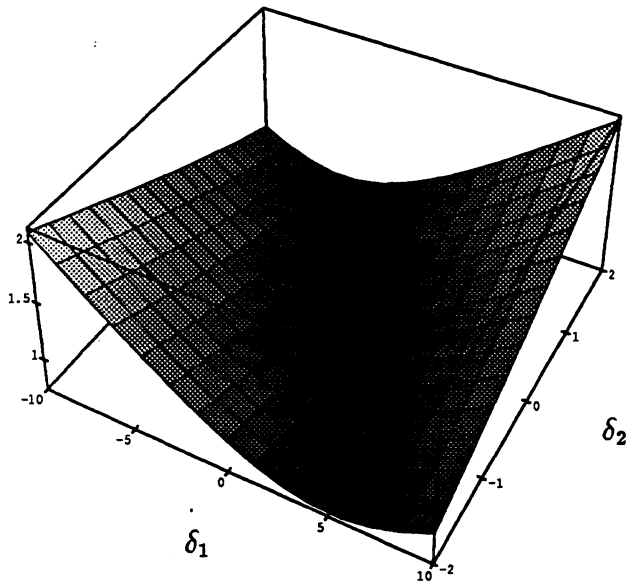

Figure 2.18: Gompertz–Makeham hazard rate. Relative risk of the narrow estimator $\sqrt{r_{\text{narr}}}/\sqrt{r_{\text{wide}}}$ plotted as a function of δ_1 and δ_2. Estimand $\mu = h(40)$. Numerical values as in the main text.

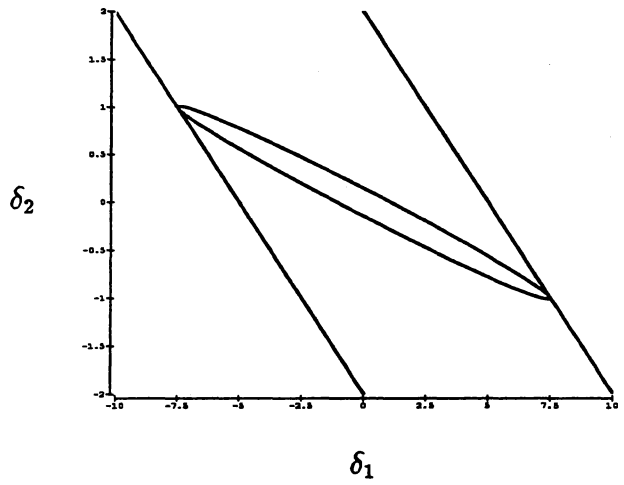

Figure 2.19: Gompertz–Makeham hazard rate. The ellipse shows the area where narrow estimation is better than wide estimation for all estimands. Note that the scale on the axes makes the ellipse appear shorter than it really is. The space between the two lines is the area where the narrow estimator is better than the wide for the estimand $\mu = h(40)$. Numerical values as in the main text.

Chapter 3

Compromise estimators

The comparison between narrow and wide estimation in the previous chapter motivates the following question. Are there estimators that do better than the wide estimator when the null model is true or almost true, and at the same time do not perform as bad as the narrow estimator when the null model is completely wrong? A natural proposal is to study estimators of the form

$$\hat{\mu}_{\text{comp}} = \mu(\hat{\xi}_{\text{comp}}),$$

where

$$\hat{\xi}_{\text{comp}} = w\hat{\xi}_{\text{wide}} + (1-w)\hat{\xi}_{\text{narr}}$$

and w is some weight function. The weight function should give increasing weight to the wide model estimator as the data indicates that the narrow model is violated. It is thus natural to let the w depend on the wide model ML-estimator of δ, $\hat{\delta} = \sqrt{n}(\hat{\gamma}_{\text{wide}} - \gamma^0)$, since this estimator estimates the deviation from the narrow model. In addition we shall allow w to depend on $\hat{\theta}$ (either the wide or the narrow version), and require it to be continuous almost everywhere in both arguments, that is $w = w(\hat{\theta}, \hat{\delta})$. Note that $\hat{\delta}$, contrary to $\hat{\gamma}_{\text{wide}}$, is not consistent, but by Lemma 2.4.2 has a limit distribution $N\{\delta, J^{22}\}$. Since w is continuous, this means that w has a limit distribution too.

Another proposal is estimators of the form

$$\hat{\mu}^*_{\text{comp}} = w\hat{\mu}_{\text{wide}} + (1-w)\hat{\mu}_{\text{narr}}.$$

Remark: One could think of more complicated versions of compromise estimators than those proposed. One could for example form convex combinations of the ML-estimators corresponding to a wide model and several different narrow sub-models. This approach could be expected to give an estimator that was better than the wide one for parameter points close to one of the narrow models, and worse than the wide estimator far from the narrow models. This idea shall not be pursued here however.

3.1 Limiting risk for the compromise estimators

Our first step is to prove that the two compromise estimators yield equivalent limit distributions. We start by considering the limit distribution of the second compromise estimator. Use a Taylor expansion of both terms around ξ^0 to obtain:

$$\sqrt{n}\left(\hat{\mu}^*_{\text{comp}} - \mu(\xi^0)\right) =$$

$$w\sqrt{n}\mu(\hat{\xi}_{\text{wide}}) + (1-w)\sqrt{n}\mu(\hat{\xi}_{\text{narr}}) - \sqrt{n}\mu(\xi^0) =$$

$$w\sqrt{n}\left[\mu(\xi^0) + \frac{\partial}{\partial\xi'}\mu(\xi^0)(\hat{\xi}_{\text{wide}} - \xi^0) + \frac{1}{2}(\hat{\xi}_{\text{wide}} - \xi^0)'\frac{\partial^2}{\partial\xi\partial\xi'}\mu(\tilde{\xi})(\hat{\xi}_{\text{wide}} - \xi^0)\right] +$$

$$(1-w)\sqrt{n}\left[\mu(\xi^0) + \frac{\partial}{\partial\xi'}\mu(\xi^0)(\hat{\xi}_{\text{narr}} - \xi^0) + \frac{1}{2}(\hat{\xi}_{\text{narr}} - \xi^0)'\frac{\partial^2}{\partial\xi\partial\xi'}\mu(\tilde{\tilde{\xi}})(\hat{\xi}_{\text{narr}} - \xi^0)\right] -$$

$$\sqrt{n}\mu(\xi^0).$$

The two remainder terms go to zero by arguments given in the proof of Lemma 2.4.4. The expression above consequently has the same limit distribution as

$$\sqrt{n}\frac{\partial}{\partial\xi'}\mu(\xi^0)\left[w\hat{\xi}_{\text{wide}} + (1-w)\hat{\xi}_{\text{narr}} - \xi^0\right] = \frac{\partial}{\partial\xi'}\mu(\xi^0)\sqrt{n}(\hat{\xi}_{\text{comp}} - \xi^0).$$

Now consider the limit distribution of the first compromise estimator:

$$\sqrt{n}\left(\mu(\hat{\xi}_{\text{comp}}) - \mu(\xi^0)\right) =$$

$$\sqrt{n}\left[\mu(\xi^0) + \frac{\partial}{\partial\xi'}\mu(\xi^0)(\hat{\xi}_{\text{comp}} - \xi^0) + \frac{1}{2}(\hat{\xi}_{\text{comp}} - \xi^0)'\frac{\partial^2}{\partial\xi\partial\xi'}\mu(\tilde{\xi})(\hat{\xi}_{\text{comp}} - \xi^0) - \mu(\xi^0)\right].$$

Again the remainder goes to zero (we shall soon prove that $\sqrt{n}(\hat{\xi}_{\text{comp}} - \xi^0)$ converges in distribution), and the limit distribution is immediately seen to be the same as that of

$$\frac{\partial}{\partial\xi'}\mu(\xi^0)\sqrt{n}(\hat{\xi}_{\text{comp}} - \xi^0),$$

implying that the two compromise estimators have the same limit distribution.

As we have now shown that the two compromise estimators are equivalent from an asymptotic point of view, we can concentrate on the first one for deriving the limit distribution. The first step will be to determine the simultaneous limit of the narrow and wide ML-estimators.

Lemma 3.1.1 *The simultaneous limit of the ML-estimators is given by*

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_{\text{wide}} - \theta \\ \hat{\gamma}_{\text{wide}} - \gamma^0 \\ \hat{\theta}_{\text{narr}} - \theta \end{pmatrix} \xrightarrow{D} \begin{pmatrix} J^{-1}\begin{pmatrix} M \\ N \end{pmatrix} + \begin{pmatrix} 0 \\ \delta \end{pmatrix} \\ J_{11}^{-1}M + J_{11}^{-1}J_{12}\delta \end{pmatrix},$$

where

$$\begin{pmatrix} M \\ N \end{pmatrix} \sim N\{0, J\}.$$

Proof: This result generalizes Lemma 2.4.2 and 2.4.3. Recall from the proofs of the two lemmas that

$$b_{n,\text{wide}} \xrightarrow{D} N\{\begin{pmatrix} J_{12} \\ J_{22} \end{pmatrix}\delta, J\}$$

and

$$b_{n,\text{narr}} \xrightarrow{D} N\{J_{12}\delta, J_{11}\},$$

which we write

$$\begin{pmatrix} b_{n,\text{wide}} \\ b_{n,\text{narr}} \end{pmatrix} \xrightarrow{D} \begin{pmatrix} M \\ N \\ M \end{pmatrix} + \begin{pmatrix} J_{12} \\ J_{22} \\ J_{12} \end{pmatrix}\delta,$$

where M and N are distributed as in the statement of the lemma. That the convergence really is simultaneous follows from the simple fact that $b_{n,\text{narr}}$ is a sub-vector of $b_{n,\text{wide}}$. (The definition of convergence in distribution requires the expectation of a bounded, continuous function of $b_{n,\text{wide}}$ and $b_{n,\text{narr}}$ to converge to the expectation of the same function in the limit distribution. But since a function of $b_{n,\text{wide}}$ and $b_{n,\text{narr}}$ can always be considered as a function of $b_{n,\text{wide}}$ alone, the simultaneous convergence follows from the convergence of $b_{n,\text{wide}}$.)

Now combine the two equations from the proofs of Lemma 2.4.2 and 2.4.3 in one equation:

$$\begin{pmatrix} A_{n,\text{wide}} & 0 \\ 0 & A_{n,\text{narr}} \end{pmatrix}\sqrt{n}\begin{pmatrix} \hat{\theta}_{\text{wide}} - \theta \\ \hat{\gamma}_{\text{wide}} - \gamma^0 \\ \hat{\theta}_{\text{narr}} - \theta \end{pmatrix} = \begin{pmatrix} b_{n,\text{wide}} \\ b_{n,\text{narr}} \end{pmatrix}.$$

Recall that the limits in probability of $A_{n,\text{wide}}$ and $A_{n,\text{narr}}$ is respectively J and J_{11}, and the result follows from Lemma 2.2.4. □

The next step is to derive the risk matrix of $\hat{\xi}_{\text{comp}}$:

Lemma 3.1.2 *The risk matrix of the compromise estimator of* ξ, $\hat{\xi}_{\text{comp}} = w\hat{\xi}_{\text{wide}} + (1 - w)\hat{\xi}_{\text{narr}}$, *is given by:*

$$R_{\text{comp}}(\theta, \delta) = \begin{pmatrix} J_{11}^{-1} + J_{11}^{-1}J_{12}R(\delta)J_{21}J_{11}^{-1} & -J_{11}^{-1}J_{12}R(\delta) \\ -R(\delta)J_{21}J_{11}^{-1} & R(\delta) \end{pmatrix},$$

where $R(\delta)$ *is the risk matrix for an estimator of* δ *of the form* $w(\theta, Z)Z$, *and* $Z \sim N\{\delta, J^{22}\}$.

Proof: By Lemma 3.1.1 and the continuous mapping theorem (since w is almost everywhere continuous) and a generalized version of the Cramér rules

$$\hat{\xi}_{\text{comp}} \xrightarrow{D} w(\theta, J^{21}M + J^{22}N + \delta)J^{-1}\begin{pmatrix} M \\ N \end{pmatrix} + \Big(1 - w(\theta, J^{21}M + J^{22}N + \delta)\Big)\begin{pmatrix} J_{11}^{-1}M + J_{11}^{-1}J_{12}\delta \\ -\delta \end{pmatrix}.$$

Denote this limit variable by L. In order to obtain a simple expression for the risk matrix, define:

$$Z = J^{21}M + J^{22}N + \delta.$$

Observe that $Z \sim N\{\delta, J^{22}\}$:

$$\text{Var}\, Z = \text{Var}\,\left(J^{21}, J^{22}\right)\begin{pmatrix} M \\ N \end{pmatrix} = \left(J^{21}, J^{22}\right)J\begin{pmatrix} J^{12} \\ J^{22} \end{pmatrix} = (0, I)\begin{pmatrix} J^{12} \\ J^{22} \end{pmatrix} = J^{22}.$$

Note further that Z and M are independent, since their simultaneous distribution is multinormal and their covariance is zero:

$$\text{Cov}(Z, M') = \text{Cov}[\left(J^{21}, J^{22}\right)\begin{pmatrix} M \\ N \end{pmatrix}, M'] = \left(J^{21}, J^{22}\right)\begin{pmatrix} J_{11} \\ J_{21} \end{pmatrix} = 0.$$

Now rewrite the expression for L in terms of Z and M:

$$L = w(\theta, Z)\begin{pmatrix} J^{11}M + J^{12}(J^{22})^{-1}(Z - J^{21}M - \delta) \\ Z - \delta \end{pmatrix} + \Big(1 - w(\theta, Z)\Big)\begin{pmatrix} J_{11}^{-1}M + J_{11}^{-1}J_{12}\delta \\ -\delta \end{pmatrix}.$$

Recall from (2.6) and (2.10) on p. 16 that $J^{11} = J_{11}^{-1} - J_{11}^{-1}J_{12}J^{21}$ and $J^{12}(J^{22})^{-1} = -J_{11}^{-1}J_{12}$.

This simplifies the first vector, and the whole expression reduces to

$$L = \begin{pmatrix} J_{11}^{-1}M + J_{11}^{-1}J_{12}\big(\delta - w(\theta, Z)Z\big) \\ w(\theta, Z)Z - \delta \end{pmatrix}. \tag{3.1}$$

The conclusion of the lemma now follows easily by calculating $R_{\text{comp}}(\theta, \delta) = ELL' = EE(LL' \mid Z)$. □

Note that this lemma shows that it does not matter for the large sample risk of the compromise estimator whether the weight function depends on θ or $\hat{\theta}$. The weight functions $w(\hat{\theta}, \delta)$ and $w(\theta, \hat{\delta})$ give exactly the same limit distributions. When discussing specific weight functions later on, we shall let the weight functions depend on θ with the understanding that in practice θ should be replaced by a consistent estimator.

We now call upon Lemma 2.4.4 to obtain the main result of this section:

Theorem 3.1.1 *The risk function of the two compromise estimators of μ, $\mu(\hat{\xi}_{\text{comp}})$ and $\hat{\mu}^*_{\text{comp}} = w\hat{\mu}_{\text{wide}} + (1-w)\hat{\mu}_{\text{narr}}$, is given by*

$$r_{\text{comp}}(\theta, \delta) = \frac{\partial \mu}{\partial \xi'} \begin{pmatrix} J_{11}^{-1} + J_{11}^{-1}J_{12}R(\delta)J_{21}J_{11}^{-1} & -J_{11}^{-1}J_{12}R(\delta) \\ -R(\delta)J_{21}J_{11}^{-1} & R(\delta) \end{pmatrix} \frac{\partial \mu}{\partial \xi}.$$

The risk function can also be written as

$$r_{\text{comp}}(\theta, \delta) = b'R(\delta)b + \tau_0^2,$$

with b and τ_0^2 defined as in Theorem 2.4.1. As in the preceding lemma, $R(\delta)$ is the the risk matrix for an estimator of δ of the form $w(\theta, Z)Z$, where $Z \sim N\{\delta, J^{22}\}$.

We observe that setting $w \equiv 1$ or $w \equiv 0$ immediately gives back the results of Theorem 2.4.1 and 2.4.2.

In the case of a one-dimensional γ, this result simplifies to the theorem on p. 15 of EMMM.

After seeing this theorem, a natural first question is to ask whether there exist estimators that are uniformly better than $\hat{\mu}_{\text{wide}}$. Somebody familiar with the "Stein phenomenon" or "shrinkage estimators" might believe this to be the case if the dimension of δ is greater than two. It is a well known result in this field that when estimating a multinormal mean, the standard estimator Z is inadmissible if the dimension is 3 or more. Unfortunately for us, this inadmissibility refers to the sum of squared errors loss[1], which is only the trace of the risk matrix $R(\delta)$. The fact that there exist estimators that make the trace of $R(\delta)$ uniformly smaller than the standard estimator Z, is not enough to conclude that there exist estimators that uniformly improve $b'R(\delta)b$. Actually, our next step is to prove the opposite: There does not exist any estimator of δ, based on Z, that makes $b'R(\delta)b$ uniformly smaller

[1] Actually, the result can be proved for a number of similar losses.

than the standard estimator Z does. This in turn immediately implies the first part of the following theorem:

Theorem 3.1.2

(i) For any given θ, there exist no estimators in the class of compromise estimators that improve upon the wide model estimator $\hat{\mu}_{\text{wide}}$ for all δ.

(ii) The narrow model estimator $\hat{\mu}_{\text{narr}}$ is also admissible in the above sense.

(iii) The wide model estimator is the unique (a.e.) minimax estimator in the class of compromise estimators. This property is also valid for any fixed θ with $b \neq 0$.

(iv) The wide model estimator is the only (asymptotically) unbiased estimator in the class of compromise estimators.

Proof: Note that

$$\begin{aligned} b'R(\delta)b &= b'E\big(w(Z)Z-\delta\big)\big(w(Z)Z-\delta\big)'b \\ &= E\big(w(Z)b'Z-b'\delta\big)^2. \end{aligned} \tag{3.2}$$

(We allow ourselves to write the weight function as $w(Z)$, thus not making the dependence on θ explicit.) If $b = 0$, all estimators have the same risk, and the statements (i) and (ii) are obvious. Suppose now that $b \neq 0$. Write the risk function in (3.2) as

$$r(\delta, d) = E(d(Z) - b'\delta)^2,$$

now visualizing the dependence on the estimator d. If we can show that $b'Z$ and 0 are admissible for $b'\delta$ in this situation (θ considered known), then (i) and (ii) will clearly follow.

Now let B be an orthogonal matrix with the first line equal to $b'(J^{22})^{1/2}$. (Remember that $Z \sim N\{\delta, J^{22}\}$.) Let $A = B(J^{22})^{-1/2}$. Clearly A will be invertible. Define the transformed variable $V = AZ$. Since this is a one-to-one transformation, we could just as well base the estimation on V. The distribution of V is $N\{\zeta, \Sigma\}$, where

$$\zeta = A\delta = \begin{pmatrix} b'\delta \\ \zeta_2 \\ \vdots \\ \zeta_r \end{pmatrix} = \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_r \end{pmatrix}$$

and

$$\Sigma = AJ^{22}A' = B(J^{22})^{-1/2}J^{22}(J^{22})^{-1/2}B' = BB'.$$

Since B is orthogonal, Σ must be diagonal, implying the components of V to be independent. Note also that Σ only depends on θ and can thus be considered known in the present situation. Our estimation problem is now, rewritten in terms of V, to minimize:

$$r(\delta, d^*) = E(d^*(V) - \zeta_1)^2.$$

That is we are estimating the expectation of the first component in a vector of independent normals. Furthermore the expectations of the components can vary freely over $\mathcal{R}^q$ since A is invertible. We would thus expect the other components of V to be of no use in the estimation. This is true; an estimator based on V_1 that is admissible when only V_1 is present, will continue to be admissible when the whole vector is present. A proof can be found in Lehmann (1983) Lemma 4.3.2 p. 269.

We can now use the well known fact that both the observation itself and the "silly" estimator 0 is admissible when estimating the mean of a one-dimensional normal variable. Thus V_1 and 0 are admissible for ζ_1. This implies in terms of Z, that both $b'Z$ and 0 are admissible for $b'\delta$. (Note that $V_1 = (AZ)_1 = b'(J^{22})^{1/2}(J^{22})^{-1/2}Z = b'Z$.) This completes the proof of (i) and (ii).

The minimax property of the wide estimator can similarly be deduced from the same property of Z_1 in the normal situation. We shall give an alternative argument, however.

Consider $b'Z$ as an estimator of $b'\delta$, in the situation with θ known. As shown above this estimator is admissible. Further, the estimator has constant risk $b'J^{22}b$. By Lemma 4.3.3 p. 276 of Lehmann (1983), this implies that the estimator is minimax. Suppose now that there exists another minimax estimator. In that case the two estimators must have exactly the same risk, otherwise the admissibility of $b'Z$ would be violated. But by Theorem 1.6.5 p. 52 of Lehmann (1983), two estimators with the same risk function must be equal with probability one, thus proving uniqueness. By (3.2) the minimax properties of $b'Z$ implies corresponding properties for the wide estimator in the class of compromise estimators, thus completing the proof of (iii).

Now finally turn to the proof of (iv). Note first that by the last part of Lemma 2.4.4 the compromise estimator of μ will be (asymptotically) unbiased if and only if the corresponding compromise estimator of ξ is. To require unbiasedness of the ξ estimator is equivalent to require the expectation of the limit variable L given by (3.1) on p. 56 to be zero. This is again clearly equivalent to require the expectation of $w(\theta, Z)Z$ to be equal to δ. Using the fact that Z is the only unbiased estimator of the mean in the multinormal distribution, it follows that $w(Z) = 1$ is the only weight function achieving unbiasedness. □

We have shown that there exists no overall champion among the compromise estimators, but we could of course still achieve our aim of constructing estimators that behave better than the wide estimator near the null model and better than the narrow estimator far from the null model. The question is how to choose the weight function w. Remember that δ is the normed distance from the null model and that Z estimates this quantity. A natural proposal is thus to let the weight function depend on estimated weighted distance from the

null model:

$$w(Z) = g(Z'WZ)$$

for some weight matrix W and function g. To choose the weight matrix, recall that the risk of the narrow estimator can be given by $(b'\delta)^2 + \tau_0^2$. The term $(b'\delta)^2$ could be said to measure the increased risk of the narrow estimator due to the incorrectness of the narrow model. If we chose the weight matrix equal to bb', then $Z'WZ = (b'Z)^2$ would estimate exactly this quantity. We shall be slightly more general and decide to consider weight functions of the form

$$w(Z) = g(b'Z)$$

for some function g. The risk function of the compromise estimator is then given by

$$R_{\text{comp}}(\theta, \delta) = E\Big(g(b'Z)b'Z - b'\delta\Big)^2 + \tau_0^2.$$

Define now a new variable V by

$$V = (b'J^{22}b)^{-1/2}b'Z.$$

Clearly V will be distributed as $N(a, 1)$, where

$$a = (b'J^{22}b)^{-1/2}b'\delta.$$

In particular if γ, and thus b, is one-dimensional, a simplifies to $\delta/\sqrt{J^{22}}$. In terms of V the risk function becomes:

$$\begin{aligned} R_{\text{comp}}(\theta, \delta) &= E\Big[g\Big((b'J^{22}b)^{1/2}V\Big)(b'J^{22}b)^{1/2}V - (b'J^{22}b)^{1/2}a\Big]^2 + \tau_0^2 \\ &= b'J^{22}bE\Big(h(V) - a\Big)^2 + \tau_0^2. \end{aligned} \tag{3.3}$$

where

$$h(V) = g\Big((b'J^{22}b)^{1/2}V\Big).$$

The expectation in (3.3) is simply the risk under squared error loss for an estimator of a in the one observation $N(a, 1)$ situation. Note that (3.3) has the same form as the wide risk function with the first term modified by the $N(a, 1)$ risk.

This could be used to compute the risk function of the compromise estimator for a given weight function $w(Z) = g(b'Z)$. But the result could also be used the other way around: Construct any estimator $h(V)$ for a in the $N(a, 1)$ situation, and transport it to a

compromise estimator through the relation

$$w(Z) = h\left((b'J^{22}b)^{-1/2}b'Z\right). \tag{3.4}$$

A genuine estimator is obtained by replacing θ by an estimate. (The limit distribution is unaffected by the insertion of any consistent estimate for θ as shown above.)

A particular consequence of the results above is that two compromise estimators on this form could be compared by solely comparing the two corresponding $N(a,1)$ estimators. The N(a,1) risks determine for each point in the parameter space which compromise estimator that has the better risk, but will not reflect the relative magnitude of the two risks. We shall therefore concentrate on the full risk, which depend on b and τ_0^2 as well as the $N(a,1)$ risk.

3.2 Examples

In this section we shall give a couple of natural compromise estimators in the $N(a,1)$ situation, and determine the risk of the corresponding compromise estimator in the Mild Regression example.

The $N(a,1)$ estimator we shall consider is the "pre-test" estimator, given by

$$\hat{a}_{\text{pre}} = I(Z^2 > 1)Z$$

and the "empirical Bayes", given by

$$\hat{a}_{\text{eb}} = \frac{1}{1+Z^2}Z.$$

We shall only give a brief description of these two estimators. A more complete discussion of these two and a number of others are found in EMMM Chapter 5.

The pre-test estimator corresponds to testing the hypothesis $Z \neq 0$ at a certain level (namely 31.7), and using the wide estimator if accepted and the narrow otherwise. This is a natural and presumably commonly used model choice strategy. Asymptotically it corresponds to the Akaike criterion (cf. EMMM).

The empirical Bayes estimator has received its name since it is the result of the following procedure. Give a a prior distribution of the form $N(0,\sigma^2)$, and compute the Bayes estimator. Since $Ea^2 = \sigma^2$ and Z estimates a, Z^2 could be considered as an estimate for σ^2. Substitute Z^2 for σ^2 in the expression for the Bayes estimator and the empirical Bayes estimator results.

Let us now see how these two compromise estimators behave in the Mild Regression example:

EXAMPLE 3 CONTINUED (MILD REGRESSION)

The two $N(a,1)$ estimators, can be transported to estimators in the Mild Regression example by (3.4). (Substitute for example the wide estimator for θ.) The risk functions can the be determined by (3.3), where the $N(a,1)$ risks must be computed by numerical integration.

The figure 3.1 gives a numerical comparison. Note that the estimators have crossing risk functions, thus making it necessary to somehow determine the relative importance of the different parameter points in order to choose between them. □

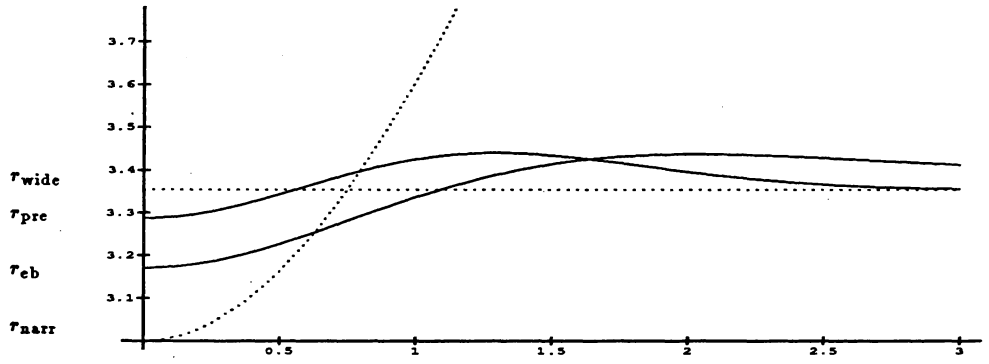

Figure 3.1: Square root of risk functions: Mild regression, estimand $\mu = \theta + 2\gamma$, ($\sigma = 3$, $\tau = 4$)

Chapter 4

Bayes estimators

In Chapter 2 we obtained a sharp criterion, giving in terms of δ the area where narrow model based ML-estimation is more precise than wide model based ML-estimation. Unfortunately, δ will of course be unknown in practice, so this will not yield an applicable criterion for model choice directly.

The compromise estimators of Chapter 3, giving data directed weights to each model, did not have this problem. But as was demonstrated in Chapter 3, there is no overall champion among the compromise estimators. Two competing estimators will typically have risk functions that cross: While one estimator is better in a certain region of the parameter space, the other is better outside this region. In my view, these facts amount to saying that an optimal choice of model or estimator must eventually depend on what is believed ápriori about the parameter.

Remark: Other optimality criteria can of course be proposed, two of the most important ones perhaps being the minimax and unbiasedness criteria. At the risk of disagreeing with a number of statisticians, I take the stand that both these criteria are quite artificial in most situations. And maybe some defenders of of these criteria are not so consequent after all, given the fact that both criteria correspond to choosing the wide model in all situations. Confer Theorem 3.1.2.

If one agrees that it is important to consider prior information, it is natural to ask for a more systematic way to use prior information than simply to pick one of the numerous compromise estimators. A natural answer to this question is to consider Bayes estimators, and that is our theme for the present chapter. By Bayes estimator we shall always mean the one with respect to squared error loss.

The discussion between "Bayesians" and "objectivists" is long and well known. I shall not repeat all the arguments here, but cannot resist the temptation to make a few additional remarks.

The most important point in favor of Bayes methodology is in my view that it incorporates prior knowledge in a straightforward, visible and systematic manner. I believe that in most situations it is impossible to make a rational choice between estimators without taking prior knowledge into consideration. (Include as prior knowledge the special case

where the entire parameter is believed to be "equally likely", typically corresponding to non-informative priors.)

For example consider a person that has read Chapter 2 of this paper, and has noted that narrow ML-estimation is more precise than wide for a range of parameter values close to the narrow model. He might with good reason decide to use narrow estimation instead of wide if he believed the correct model to be quite close to the narrow one, even though he knew that the narrow model was certainly not exactly correct!

Now it may seem quite paradoxical to deliberately choose the wrong model in order to obtain better results. [1] This paradox is resolved elegantly in a Bayesian framework. The solution would of course be to choose the correct model, namely the wide one, but to place a prior distribution on the parameter, reflecting the belief that values near the null model is considered more likely than others. This would probably result in an estimator reflecting the prior knowledge more accurately than merely choosing the narrow ML-estimator (or some compromise estimator), and as an additional feature the Bayes approach would explicitly state the prior beliefs employed.

After these general remarks, it is time to turn to the study of Bayes estimators in our asymptotic framework. In section 4.1 we shall study the estimators from the frequentist point of view, that is through the risk function. But in section 4.3 we shall compare Bayes estimators and ML-estimators on the Bayes estimator's premises. The criterion will then be the limiting Bayes risk, which we define as the expectation of the risk function in the prior distribution.

When comparing estimators with respect to Bayes risk, the Bayes estimator will of course by definition be optimal (under mild regularity in our limiting case). But from a Bayesian point of view it will be interesting for two reasons to see how much is lost by using either narrow or wide model ML-estimators. The first reason is that ML-estimation is in widespread use, whether one likes it or not. It is thus interesting to evaluate the consequences of this "refusal to take prior knowledge into consideration". The second reason is computational. The Bayes estimators are often hard to compute, and in many such situations ML-estimators are readily available. If the loss incurred by using either the wide or narrow ML-estimator is small, it may be used as an approximation to the Bayes estimator.

4.1 Limiting risk for the Bayes estimators

The aim of the present section is to derive the risk functions for the Bayes estimators. Our previous wide model will make up the conditional model given the parameters.

When constructing Bayes estimators, we could place a fixed prior distribution on $(\theta', \gamma')'$. This would lead to a Bayes estimator that is large sample equivalent to the (wide model) ML-estimator, and is thus not very interesting from our point of view. To reflect

[1] We do not of course argue against the choice of simplified models motivated by the need to obtain mathematically tractable solutions. This is an essential feature in all branches of applied mathematics, but not the theme for our discussion.

a prior opinion that will not become irrelevant as the amount of data grows, it is more natural to place a fixed prior density on $(\theta', \delta')'$. (Remember that $\gamma = \gamma^0 + \delta/\sqrt{n}$ in our chosen large sample framework.) As is shown below, the marginal distribution of θ will have no effect on the risk function (it will matter when computing the Bayes risk, though), so the important aspect of the prior density is the conditional density of δ given θ.

We shall start by determining the risk function in the special case where the conditional density of δ given θ is normal, since this is somewhat easier than the general case. Parallelling the case for the ML-estimators, we shall determine the risks for the Bayes estimator of ξ first, and then use this to determine the risk of the Bayes estimator of μ.

Lemma 4.1.1 *Suppose that the conditional prior density of δ given θ is $N\{0, K\}$, where $K = K(\theta)$ may depend on θ. The prior density of θ is assumed to be continuous, but can otherwise be arbitrary. Then the risk matrix of the Bayes estimator of ξ is given by:*

$$R_{\text{Bayes}}(\theta, \delta) = \left[J + \begin{pmatrix} 0 & 0 \\ 0 & K^{-1} \end{pmatrix}\right]^{-1} \left[J + \begin{pmatrix} 0 & 0 \\ 0 & K^{-1}\delta\delta'K^{-1} \end{pmatrix}\right] \left[J + \begin{pmatrix} 0 & 0 \\ 0 & K^{-1} \end{pmatrix}\right]^{-1}.$$

Proof: Recall that the Bayes estimator with respect to weighted squared error is the conditional expectation of the parameter given the data. The conditional distribution of ξ given data is:

$$f(\xi|\text{data}) = \frac{L(\xi) f_\xi(\xi)}{\int L(\xi) f_\xi(\xi)\, d\xi},$$

where $L(\xi)$ is the likelihood function (simultaneous density) of the data. The Bayes-estimator is then given by

$$\begin{aligned} \hat{\xi}_{\text{Bayes}} &= \frac{\int \xi L(\xi) f_\xi(\xi)\, d\xi}{\int L(\xi) f_\xi(\xi)\, d\xi} \\ &= \frac{\int \xi L(\xi) f_\theta(\theta) f_{\gamma|\theta}(\gamma, \theta)\, d\xi}{\int L(\xi) f_\theta(\theta) f_{\gamma|\theta}(\gamma, \theta)\, d\xi}. \end{aligned}$$

To start the proof, we shall use a "trick" which consists of substituting $\xi = \hat{\xi} + \frac{1}{\sqrt{n}}t$ in the two integrals. (For simplicity of notation we shall drop the subscripts on the wide model ML-estimators $\hat{\xi}_{\text{wide}}$, $\hat{\theta}_{\text{wide}}$ etc. in this section.) As a motivation for this substitution, remember that $L(\xi)$ will be increasingly concentrated around $\hat{\xi}$ as n grows. An example of a similar method of proof in a different context can be found in Hjort (1986).

The Jacobi-matrix of this substitution is $\frac{\partial \xi}{\partial t'} = \frac{1}{\sqrt{n}} I$ with determinant $\left|\frac{\partial \xi}{\partial t'}\right| = \left(\frac{1}{\sqrt{n}}\right)^r$.

The result of the substitution is thus after canceling the two Jacobi-determinants:

$$\hat{\xi}_{\text{Bayes}} = \frac{\int \left(\hat{\xi} + \frac{1}{\sqrt{n}}t\right) L(\hat{\xi} + \frac{1}{\sqrt{n}}t) f_\theta(\hat{\theta} + \frac{1}{\sqrt{n}}t_1) f_{\gamma|\theta}(\hat{\gamma} + \frac{1}{\sqrt{n}}t_2, \hat{\theta} + \frac{1}{\sqrt{n}}t_1)\, dt}{\int L(\hat{\xi} + \frac{1}{\sqrt{n}}t) f_\theta(\hat{\theta} + \frac{1}{\sqrt{n}}t_1) f_{\gamma|\theta}(\hat{\gamma} + \frac{1}{\sqrt{n}}t_2, \hat{\theta} + \frac{1}{\sqrt{n}}t_1)\, dt}.$$

(We have partitioned t into t_1 and t_2 with dimensions corresponding to θ and γ respectively.) The last equation can be written as:

$$\sqrt{n}(\hat{\xi}_{\text{Bayes}} - \hat{\xi}) = \frac{\int t \frac{L(\hat{\xi}+\frac{1}{\sqrt{n}}t)}{L(\hat{\xi})} \frac{f_\theta(\hat{\theta}+\frac{1}{\sqrt{n}}t_1)}{f_\theta(\theta)} n^{-q/2} f_{\gamma|\theta}(\hat{\gamma} + \frac{1}{\sqrt{n}}t_2, \hat{\theta} + \frac{1}{\sqrt{n}}t_1)\, dt}{\int \frac{L(\hat{\xi}+\frac{1}{\sqrt{n}}t)}{L(\hat{\xi})} \frac{f_\theta(\hat{\theta}+\frac{1}{\sqrt{n}}t_1)}{f_\theta(\theta)} n^{-q/2} f_{\gamma|\theta}(\hat{\gamma} + \frac{1}{\sqrt{n}}t_2, \hat{\theta} + \frac{1}{\sqrt{n}}t_1)\, dt}. \tag{4.1}$$

The next step is to determine the limit in distribution of the two integrands. Consider first

$$\log \frac{L(\hat{\xi} + \frac{1}{\sqrt{n}}t)}{L(\hat{\xi})}.$$

Use a Taylor expansion around $\hat{\xi}$ to obtain

$$\begin{aligned}
\log \frac{L(\hat{\xi} + \frac{1}{\sqrt{n}}t)}{L(\hat{\xi})} &= \log L(\hat{\xi}) + \frac{1}{\sqrt{n}} t \frac{\partial}{\partial \xi'} \log L(\hat{\xi}) + \frac{1}{2n} t' \frac{\partial^2}{\partial \xi \partial \xi'} \log L(\tilde{\xi}) t - \log L(\hat{\xi}) \\
&= \frac{1}{2n} t' \frac{\partial^2}{\partial \xi \partial \xi'} \log L(\tilde{\xi}) t \\
&= \frac{1}{2} t' \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \xi \partial \xi'} \log f(Y_{ni}|\tilde{\xi}) t \\
&\xrightarrow{P} -\frac{1}{2} t' J t.
\end{aligned}$$

The last convergence follows from lemmas 2.2.2 and 2.3.1. (Note that $\tilde{\xi}$ is situated somewhere on the line between $\hat{\xi}$ and $\hat{\xi} + \frac{1}{\sqrt{n}}t$. The latter two variables both converge in probability to ξ^0, implying that $\tilde{\xi}$ also must converge in probability to ξ^0.) We thus conclude by Slutsky that

$$\frac{L(\hat{\xi} + \frac{1}{\sqrt{n}}t)}{L(\hat{\xi})} \xrightarrow{P} \exp\left\{-\frac{1}{2} t' J t\right\}.$$

The limit in probability of

$$\frac{f_\theta(\hat{\theta} + \frac{1}{\sqrt{n}}t_1)}{f_\theta(\theta)}$$

is immediately seen to be 1 by Cramér and Slutsky since $f_\theta(\theta)$ is assumed continuous.

Now turn to

$$n^{-q/2} f_{\gamma|\theta}(\hat{\gamma} + \frac{1}{\sqrt{n}} t_2, \hat{\theta} + \frac{1}{\sqrt{n}} t_1). \tag{4.2}$$

Since $\gamma = \gamma^0 + \delta/\sqrt{n}$, the prior distribution for γ given θ is given by $N\{\gamma^0, \frac{1}{n}K\}$. Inserting the expression for this normal density in (4.2) we obtain:

$$\frac{\exp\left\{-\frac{1}{2}\left(\hat{\gamma} + \frac{1}{\sqrt{n}}t_2 - \gamma^0\right)' \left[\frac{1}{n}K(\hat{\theta} + \frac{1}{\sqrt{n}}t_1)\right]^{-1} \left(\hat{\gamma} + \frac{1}{\sqrt{n}}t_2 - \gamma^0\right)\right\}}{n^{q/2}(2\pi)^{q/2}\left(\det \frac{1}{n}K(\hat{\theta} + \frac{1}{\sqrt{n}}t_1)\right)^{1/2}},$$

which is equivalent to

$$\frac{\exp\left\{-\frac{1}{2}\left(t_2 + \sqrt{n}(\hat{\gamma} - \gamma^0)\right)' \left[K(\hat{\theta} + \frac{1}{\sqrt{n}}t_1)\right]^{-1} \left(t_2 + \sqrt{n}(\hat{\gamma} - \gamma^0)\right)\right\}}{(2\pi)^{q/2}\left(\det K(\hat{\theta} + \frac{1}{\sqrt{n}}t_1)\right)^{1/2}}.$$

Now let A be a variable distributed according to $N\{0, J^{-1}\}$. Partition A into $(A_1', A_2')'$ with dimensions corresponding to those of θ and γ as usual. For simplicity of notation introduce also $B = A + (0, \delta')'$ and $Z = A_2 + \delta$. (Note that the definition of Z corresponds to that used in Chapter 3.) Finally, remember that $\sqrt{n}(\hat{\gamma} - \gamma^0)$ converges in distribution to Z. It then follows (by the continuous mapping theorem and Cramér) that the limit distribution of the above expression may be given as:

$$(2\pi)^{-q/2} \det K(\theta)^{-1/2} \exp\left\{-\frac{1}{2}(t_2 + Z)' K(\theta)^{-1}(t_2 + Z)\right\}.$$

Combining these facts and applying Cramér rules we can now, for given t, determine the limit distribution of two processes forming the integrands of (4.1). By the Cramér–Wold device we can easily extend the result and show that the limit is simultaneous for all finite-dimensional distributions of the two processes. (We also include distributions formed by picking variables from both processes.) This fact makes it reasonable indeed to believe that the two integrals will converge to the integrals of the limit processes, and that the convergence will be simultaneous. From the results obtained above the limit should thus be

$$\sqrt{n}(\hat{\xi}_{\text{Bayes}} - \hat{\xi}) \xrightarrow{D} \frac{\int t \exp\left\{-\frac{1}{2}t'Jt\right\} \exp\left\{-\frac{1}{2}(t_2 + Z)'K^{-1}(t_2 + Z)\right\} dt}{\int \exp\left\{-\frac{1}{2}t'Jt\right\} \exp\left\{-\frac{1}{2}(t_2 + Z)'K^{-1}(t_2 + Z)\right\} dt}.$$

The precise verification of this "integration to the limit" depends on some additional technical details that shall not be given here. We refer to Hjort (1986) to see the kind of arguments needed.

Our next step is to obtain a simple expression for this limit. Define

$$S = \begin{pmatrix} 0 & 0 \\ 0 & K^{-1} \end{pmatrix}$$

and note that $J + S$ is positive definite, and thus invertible, since J is positive definite and S is positive semi-definite. Multiplying the same factors (constant in t) in numerator and denominator transforms the limit variable to

$$\frac{\int t(2\pi)^{-r/2}\big(\det(J+S)\big)^{1/2}\exp\left\{-\frac{1}{2}\big(t+(J+S)^{-1}SB\big)'(J+S)\big(t+(J+S)^{-1}SB\big)\right\}dt}{\int (2\pi)^{-r/2}\big(\det(J+S)\big)^{1/2}\exp\left\{-\frac{1}{2}\big(t+(J+S)^{-1}SB\big)'(J+S)\big(t+(J+S)^{-1}SB\big)\right\}dt}.$$

The integral in the numerator is now the expectation in a $N\{-(J+S)^{-1}SB, (J+S)^{-1}\}$ distribution and the numerator is the integral over the same density. We have thus reached the limit result

$$\sqrt{n}(\hat{\xi}_{\text{Bayes}} - \hat{\xi}) \xrightarrow{D} -(J+S)^{-1}SB.$$

To find the limit of $\sqrt{n}(\hat{\xi}_{\text{Bayes}} - \xi)$, we note that

$$\sqrt{n}(\hat{\xi} - \xi) \xrightarrow{D} A \tag{4.3}$$

and that our Cramér–Wold argument can be extended to show that this limit is simultaneous with those obtained previously. Consequently we have

$$\begin{aligned}
\sqrt{n}(\hat{\xi}_{\text{Bayes}} - \xi) &= \sqrt{n}(\hat{\xi}_{\text{Bayes}} - \hat{\xi}) + \sqrt{n}(\hat{\xi} - \xi) \\
&\xrightarrow{D} -(J+S)^{-1}S\left[A + \begin{pmatrix} 0 \\ \delta \end{pmatrix}\right] + A \\
&= -(J+S)^{-1}S\begin{pmatrix} 0 \\ \delta \end{pmatrix} + \big(I - (J+S)^{-1}S\big)A \\
&= -(J+S)^{-1}S\begin{pmatrix} 0 \\ \delta \end{pmatrix} + (J+S)^{-1}JA.
\end{aligned}$$

The risk function is thus given by

$$R_{\text{Bayes}}(\theta, \delta) = (J+S)^{-1}\left[J + \begin{pmatrix} 0 & 0 \\ 0 & K^{-1}\delta\delta'K^{-1} \end{pmatrix}\right](J+S)^{-1},$$

which is the statement of the lemma. □

Encouraged by our success with normal priors, we now turn to the general case. Before

giving the result, we state for easy reference a well known formula from multivariate analysis which will be needed in the proof.

Lemma 4.1.2 *Suppose that*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\{\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\}$$

then the conditional distribution of X_1 given X_2 is

$$N\{\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\}.$$

Now to the main result:

Lemma 4.1.3 *For continuous, in both θ and δ, but otherwise arbitrary prior density, the risk matrix of the Bayes estimator of ξ is given by:*

$$R_{\text{Bayes}}(\theta, \delta) = \begin{pmatrix} J_{11}^{-1} + J_{11}^{-1}J_{12}R(\delta)J_{21}J_{11}^{-1} & -J_{11}^{-1}J_{12}R(\delta) \\ -R(\delta)J_{21}J_{11}^{-1} & R(\delta) \end{pmatrix},$$

where $R(\delta)$ is the risk matrix of the Bayes estimator in the situation with one observation $Z \sim N\{\delta, J^{22}\}$, θ considered known and $f_{\delta|\theta}(\delta)$ as prior distribution for δ.

Proof: Remember from the proof of the preceding lemma that

$$\sqrt{n}(\hat{\xi}_{\text{Bayes}} - \hat{\xi}) = \frac{\int t \frac{L(\hat{\xi}+\frac{1}{\sqrt{n}}t)}{L(\hat{\xi})} \frac{f_\theta(\hat{\theta}+\frac{1}{\sqrt{n}}t_1)}{f_\theta(\theta)} n^{-q/2} f_{\gamma|\theta}(\hat{\gamma} + \frac{1}{\sqrt{n}}t_2, \hat{\theta} + \frac{1}{\sqrt{n}}t_1)\, dt}{\int \frac{L(\hat{\xi}+\frac{1}{\sqrt{n}}t)}{L(\hat{\xi})} \frac{f_\theta(\hat{\theta}+\frac{1}{\sqrt{n}}t_1)}{f_\theta(\theta)} n^{-q/2} f_{\gamma|\theta}(\hat{\gamma} + \frac{1}{\sqrt{n}}t_2, \hat{\theta} + \frac{1}{\sqrt{n}}t_1)\, dt}.$$

We thus need to determine the limit of

$$n^{-q/2} f_{\gamma|\theta}(\hat{\gamma} + \frac{1}{\sqrt{n}}t_2, \hat{\theta} + \frac{1}{\sqrt{n}}t_1). \tag{4.4}$$

Since $\gamma = \gamma^0 + \delta/\sqrt{n}$, the transformation theorem yields for the conditional density of γ given θ:

$$f_{\gamma|\theta}(\gamma, \theta) = |\frac{\partial \delta}{\partial \gamma'}| f_{\delta|\theta}(\delta, \theta) = n^{q/2} f_{\delta|\theta}(\sqrt{n}(\gamma - \gamma^0), \theta).$$

After substituting this in (4.4), the continuous mapping theorem immediately gives the limit (with Z as before):

$$f_{\delta|\theta}(Z + t_2, \theta).$$

Using otherwise the same arguments as in the preceding proof, we conclude that

$$\sqrt{n}(\hat{\xi}_{\text{Bayes}} - \hat{\xi}) \xrightarrow{D} \frac{\int t \exp\left\{-\frac{1}{2}t'Jt\right\} f_{\delta|\theta}(Z + t_2, \theta)\, dt}{\int \exp\left\{-\frac{1}{2}t'Jt\right\} f_{\delta|\theta}(Z + t_2, \theta)\, dt}.$$

Now single out the part of the integration concerning t_1 in the above limit:

$$\frac{\iint \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \exp\left\{-\frac{1}{2}t'Jt\right\} dt_1\, f_{\delta|\theta}(Z + t_2, \theta)\, dt_2}{\iint \exp\left\{-\frac{1}{2}t'Jt\right\} dt_1\, f_{\delta|\theta}(Z + t_2, \theta)\, dt_2}. \tag{4.5}$$

We want to evaluate the two inner integrals. As before let $(A_1', A_2')'$ have a $N\{0, J^{-1}\}$ distribution. The marginal density of A_2 can of course be calculated by an integral over the simultaneous density:

$$f_{A_2}(t_2) = \int c \exp\left\{-\frac{1}{2}t'Jt\right\} dt_1.$$

(The constants of the densities will be seen to cancel out, so we do not bother to display them explicitly.) Combining this with the other well known expression for the marginal density in the normal distribution, we obtain

$$\int \exp\left\{-\frac{1}{2}t'Jt\right\} dt_1 = \frac{c_2}{c} \exp\left\{-\frac{1}{2}t_2'(J^{22})^{-1}t_2\right\}. \tag{4.6}$$

Furthermore, the conditional expectation of A_1 given A_2 can be given by

$$E(A_1|A_2 = t_2) = \int t_1 f_{A_1|A_2}(t_1|t_2)\, dt_1 = \frac{\int t_1 c \exp\left\{\frac{1}{2}t'Jt\right\} dt_1}{c_2 \exp\left\{\frac{1}{2}t_2'(J^{22})^{-1}t_2\right\}}.$$

This conditional expectation is also given by Lemma 4.1.2 and we obtain:

$$\int t_1 \exp\left\{\frac{1}{2}t'Jt\right\} dt_1 = \frac{c_2}{c} J^{12}(J^{22})^{-1}t_2 \exp\left\{\frac{1}{2}t_2'(J^{22})^{-1}t_2\right\}. \tag{4.7}$$

After calculating the inner integrals by (4.6) and (4.7), (4.5) simplifies to:

$$\frac{\int \begin{pmatrix} J^{12}(J^{22})^{-1}t_2 \\ t_2 \end{pmatrix} \exp\left\{-\frac{1}{2}t_2'(J^{22})^{-1}t_2\right\} f_{\delta|\theta}(Z + t_2, \theta)\, dt_2}{\int \exp\left\{-\frac{1}{2}t_2'(J^{22})^{-1}t_2\right\} f_{\delta|\theta}(Z + t_2, \theta)\, dt_2}.$$

Now substitute $t_2 = \delta - Z$ in both integrals and rearrange a little to obtain:

$$\begin{pmatrix} J^{12}(J^{22})^{-1} \\ I \end{pmatrix} \frac{\int (\delta - Z) \exp\left\{-\frac{1}{2}(Z-\delta)'(J^{22})^{-1}(Z-\delta)\right\} f_{\delta|\theta}(\delta,\theta)\, d\delta}{\int \exp\left\{-\frac{1}{2}(Z-\delta)'(J^{22})^{-1}(Z-\delta)\right\} f_{\delta|\theta}(\delta,\theta)\, d\delta}.$$

The ratio in the above expression is seen to be nothing but $E(\delta|Z) - Z$, where δ is distributed as in the statement of the lemma and θ is considered known. Thus we have shown

$$\sqrt{n}(\hat{\xi}_{\text{Bayes}} - \hat{\xi}) \xrightarrow{D} \begin{pmatrix} J^{12}(J^{22})^{-1} \\ I \end{pmatrix} \big(E(\delta|Z) - Z\big). \tag{4.8}$$

Recall from (4.3) that

$$\sqrt{n}(\hat{\xi} - \xi) \xrightarrow{D} A = \begin{pmatrix} A_1 \\ Z - \delta \end{pmatrix},$$

and that this limit is simultaneous with the one obtained for $\sqrt{n}(\hat{\xi}_{\text{Bayes}} - \hat{\xi})$. We are now finally in a position to determine the desired limit variable:

$$\sqrt{n}(\hat{\xi}_{\text{Bayes}} - \xi) = \sqrt{n}(\hat{\xi}_{\text{Bayes}} - \hat{\xi}) + \sqrt{n}(\hat{\xi} - \xi) \xrightarrow{D} \begin{pmatrix} J^{12}(J^{22})^{-1} \\ I \end{pmatrix} \big(E(\delta|Z) - Z\big) + \begin{pmatrix} A_1 \\ Z - \delta \end{pmatrix}.$$

Denote the limit by L. A little rearranging and substitution of (2.10) on p. 16 yields:

$$L = \begin{pmatrix} -J_{11}^{-1}J_{12}\big(E(\delta|Z) - \delta\big) + A_1 + J_{11}^{-1}J_{12}(Z - \delta) \\ E(\delta|Z) - \delta \end{pmatrix}.$$

To compute the risk matrix, note that by Lemma 4.1.2 the conditional distribution of A_1 given Z is $N\{J^{12}(J^{22})^{-1}(Z-\delta), J^{11} - J^{12}(J^{22})^{-1}J^{21}\}$, which by (2.10) and (2.11) is equivalent to $N\{J_{11}^{-1}J_{12}(Z-\delta), J_{11}^{-1}\}$. Using this, the risk matrix can now be computed by

$$ELL' = EE(LL'|Z).$$

After computing the inner expectation, a number of terms cancel and the statement of the lemma is obtained. □

We can now use this general result to once more obtain an expression for the risk matrix in case of a normal prior: Let δ have a $N\{0, K\}$ density, and let Z be distributed as $N\{\delta, J^{22}\}$ as usual. The Bayes estimator of δ based on Z is in this situation well known to be (cf. Berger (1985) p. 140)

$$\hat{\delta}_{\text{Bayes}} = \left[(J^{22})^{-1} + K^{-1}\right]^{-1} (J^{22})^{-1} Z.$$

The risk matrix of this estimator is

$$\begin{aligned} R(\delta) &= E(\hat{\delta}_{\text{Bayes}} - \delta)(\hat{\delta}_{\text{Bayes}} - \delta)' \\ &= \text{Var}\,\hat{\delta}_{\text{Bayes}} + \left(E(\hat{\delta}_{\text{Bayes}}) - \delta\right)\left(E(\hat{\delta}_{\text{Bayes}}) - \delta\right)', \end{aligned}$$

which by an easy calculation is equivalent to

$$R(\delta) = \left[(J^{22})^{-1} + K^{-1}\right]^{-1}\left[(J^{22})^{-1} + K^{-1}\delta\delta' K^{-1}\right]\left[(J^{22})^{-1} + K^{-1}\right]^{-1}. \qquad (4.9)$$

Substituting this expression in Lemma 4.1.3 yields a new expression for the risk matrix in case of a normal prior, which the reader may want to compare to Lemma 4.1.1. (Lemma 4.1.1 could of course now have been dispensed with, but the direct proof has been included anyway for two reasons: Mainly because the proof is somewhat easier than the general version in Lemma 4.1.3, but also because it delivers an alternative formula which is not straightforward to verify directly.)

Our natural next step is now to determine the risk function of the Bayes estimator of μ. Unfortunately we cannot apply Lemma 2.4.4 since $\hat{\mu}_{\text{Bayes}}$ is *not* given by $\mu(\hat{\xi}_{\text{Bayes}})$. This is not a serious problem, however, because it is possible to establish an analogous result:

Lemma 4.1.4 *The risk function of the Bayes estimator of μ is given by*

$$r_{\text{Bayes}}(\theta, \delta) = \frac{\partial \mu}{\partial \xi'} R_{\text{Bayes}}(\theta, \delta) \frac{\partial \mu}{\partial \xi},$$

where the partial derivatives are computed at the null point ξ^0.

Proof: By definition of the Bayes estimator

$$\sqrt{n}(\hat{\mu}_{\text{Bayes}} - \mu) = \sqrt{n}\left[\int \mu(\xi^1) f(\xi^1|\text{data})\, d\xi^1 - \mu(\xi)\right].$$

Now Taylor-expand both μ functions around ξ^0 to obtain:

$$\begin{aligned} \sqrt{n}\ \Big[\int \Big[\mu(\xi^0) + \frac{\partial}{\partial \xi'}\mu(\xi^0)(\xi^1 - \xi^0) + \frac{1}{2}(\xi^1 - \xi^0)' \frac{\partial^2}{\partial \xi \partial \xi'}\mu(\tilde{\xi})(\xi^1 - \xi^0)\Big] f(\xi^1|\text{data})\, d\xi^1 \\ -\mu(\xi^0) - \frac{\partial}{\partial \xi'}\mu(\xi^0)(\xi - \xi^0) - \frac{1}{2}(\xi - \xi^0)' \frac{\partial^2}{\partial \xi \partial \xi'}\mu(\tilde{\tilde{\xi}})(\xi - \xi^0)\Big]. \end{aligned}$$

Given the right regularity conditions, both remainder terms will go to zero in probability,

and after canceling a few terms we are left with:

$$\sqrt{n}(\hat{\mu}_{\text{Bayes}} - \mu) \doteq \frac{\partial}{\partial \xi'}\mu(\xi^0)\sqrt{n}\left[\int \xi^1 f(\xi^1|\text{data})\, d\xi^1 - \xi\right] = \frac{\partial}{\partial \xi'}\mu(\xi^0)\sqrt{n}\left[\hat{\xi}_{\text{Bayes}} - \xi\right].$$

This immediately implies the statement of the lemma. □

After this preliminary we can now state the main results of this section:

Theorem 4.1.1 *Under the conditions of Lemma 4.1.3, the Bayes estimator of μ has risk function given by*

$$r_{\text{Bayes}}(\theta, \delta) = \frac{\partial \mu}{\partial \xi'}\begin{pmatrix} J_{11}^{-1} + J_{11}^{-1}J_{12}R(\delta)J_{21}J_{11}^{-1} & J_{11}^{-1}J_{12}R(\delta) \\ R(\delta)J_{21}J_{11}^{-1} & R(\delta) \end{pmatrix}\frac{\partial \mu}{\partial \xi},$$

with $R(\delta)$ as in Lemma 4.1.3. The risk function can be written as

$$r_{\text{Bayes}}(\theta, \delta) = b'R(\delta)b + \tau_0^2,$$

with b and τ_0^2 as in Theorem 2.4.1.

As a simple consequence of this theorem and (4.9), we obtain in the case of a normal prior for δ:

Corollary 4.1.1 *Suppose that the conditional prior density of δ given θ is $N\{0, K\}$, as in Lemma 4.1.1. Then the risk function of the Bayes estimator of μ is given by:*

$$r_{\text{Bayes}}(\theta, \delta) = b'\left[(J^{22})^{-1} + K^{-1}\right]^{-1}\left[(J^{22})^{-1} + K^{-1}\delta\delta'K^{-1}\right]\left[(J^{22})^{-1} + K^{-1}\right]^{-1} b + \tau_0^2.$$

Note that the risk of the wide and narrow ML-estimator is the limiting cases when respectively K^{-1} or K approach zero in some suitable fashion. Let us for example set $K = kK_0$ for some fixed K_0 and let k go to infinity. Then

$$r_{\text{Bayes}}(\theta, \delta) =$$

$$b'\left[(J^{22})^{-1} + \frac{1}{k}K_0^{-1}\right]^{-1}\left[(J^{22})^{-1} + \frac{1}{k^2}K_0^{-1}\delta\delta'K_0^{-1}\right]\left[(J^{22})^{-1} + \frac{1}{k}K_0^{-1}\right]^{-1} b + \tau_0^2 \underset{k\to\infty}{\longrightarrow}$$

$$b'J^{22}b + \tau_0^2 = r_{\text{wide}}(\theta, \delta).$$

Similarly

$$r_{\text{Bayes}}(\theta, \delta) \underset{k\to 0}{\longrightarrow} r_{\text{narr}}(\theta, \delta).$$

We thus have the following two intuitively reasonable conclusions:

- In situations with "vague" prior information about the correctness of the narrow model, corresponding to large variance in the prior distribution of δ, we could just as well use the wide ML-estimator as the Bayes estimator.
- In situations with precise prior information about the correctness of the narrow model, corresponding to a prior distribution for δ concentrated around zero, we could just as well use the narrow ML-estimator as the Bayes estimator.

4.2 Regression generalization

We want to generalize our results about Bayes estimators to regression type situations. We follow the same approach as in section 2.5. The regression covariates x_i are considered to be i.i.d. random variables with a density $f(x|\zeta)$. Now let ζ have a prior density $f(\zeta)$. We further assume that the prior information about ζ is independent of the prior information about ξ. That is, the simultaneous prior density of $(\zeta', \xi')'$ is $f(\xi)f(\zeta)$. This assumption ensure that the Bayes estimators are identical with those derived under the fixed x_i approach by the following simple argument:

$$\begin{aligned}\hat{\mu}_{\text{Bayes}} = E(\mu|\text{data}) &= \frac{\int \mu(\xi) \prod_{i=1}^n f(y_i|\xi, x_i) f(x_i|\zeta) f(\xi) f(\zeta)\, d\xi}{\int \prod_{i=1}^n f(y_i|\xi, x_i) f(x_i|\zeta) f(\xi) f(\zeta)\, d\xi} \\ &= \frac{\int \mu(\xi) \prod_{i=1}^n f(y_i|\xi, x_i) f(\xi)\, d\xi}{\int \prod_{i=1}^n f(y_i|\xi, x_i) f(\xi)\, d\xi}.\end{aligned}$$

The theory of the preceding section can now be retained unchanged by simply identifying the pair $(Y', x')'$ with our previous Y'. When computing risk functions the only visible change will be that J is computed as the expectation in the simultaneous distribution of Y and x.

Remark: As in section 2.5 there is a small point to consider here too, since ξ is not the entire parameter vector of the model any longer. We should show that this does not affect the risk formulae. If ζ is considered known, the structure of the model is identical to the non-regression case, and the old formulae remain valid. But we have shown that the form of the Bayes estimator is not affected by the prior information about ζ. And the risk function is by definition ignorant of any prior information. All our results will thus remain valid in the more realistic case of unknown ζ with a prior distribution. Cf. the remark on p. 25.

4.3 Bayes risk for the estimators

We will now give expressions for the Bayes risk of our estimators. The results will be immediate consequences of the corresponding results about risk functions. We remind the reader that we have defined (limiting) Bayes risk as the expectation of the risk function in the prior distribution. The Bayes risk will be denoted by "br".

Remark 1: Another natural definition would have been to define the limiting Bayes risk as the limit of n times the finite Bayes risk: $\mathrm{br}(\theta, \delta) = \lim_{n\to\infty} nE(\hat{\mu} - \mu)^2$ As it turns out, the two criteria are equivalent in most situations, and we choose the one giving less technical difficulties. (See the remark on p. 9.)

Remark 2: There is also a more fundamental objection to our choice of risk criterion. To be faithful to the conditional view, we should use the conditional Bayes risk given the value of the data, commonly denoted by áposteriori expected loss. In my opinion this is the more natural criterion. It is a waste of information to average over all possible data when it is already known which one occurred. (Averaging over data which are already known not to occur is of course fundamental in the frequentist perspective. See the remark on p. 26.)

Unfortunately, the conditional Bayes risk can not be assessed in our large sample framework, and we shall be satisfied to study the Bayes risk as defined above. As is well known the two criteria give the same optimal estimators, but the risk of the estimators will of course differ.

Before the data is collected the (unconditional) Bayes risk would be the natural criterion.

Corollary 4.3.1 *Under the conditions of Lemma 4.1.3 the Bayes estimator of μ has Bayes risk given by*

$$\mathrm{br}_{\mathrm{Bayes}} = Eb'R(\delta)b + E\tau_0^2.$$

The expectation should be computed under the prior distribution of θ and δ.

Corollary 4.3.2 *Suppose that the conditional prior density of δ given θ is $N\{0, K\}$, as in Lemma 4.1.1. The Bayes risk of the Bayes estimator of μ is then given by:*

$$\mathrm{br}_{\mathrm{Bayes}} = Eb'\left[(J^{22})^{-1} + K^{-1}\right]^{-1} b + E\tau_0^2.$$

Proof: Compute the expectation of the risk function from Corollary 4.1.1 by first conditioning on θ. □

Corollary 4.3.3 *Under the conditions of Lemma 4.1.3 the wide ML-estimator of μ has Bayes risk given by*

$$\mathrm{br}_{\mathrm{wide}} = Eb'J^{22}b + E\tau_0^2.$$

Corollary 4.3.4 *Under the conditions of Lemma 4.1.3 the narrow ML-estimator of μ has Bayes risk given by*

$$\mathrm{br}_{\mathrm{narr}} = Eb'\delta\delta'b + E\tau_0^2.$$

Corollary 4.3.5 *Suppose that the conditional prior density of δ given θ is $N\{0, K\}$, as in Lemma 4.1.1. The Bayes risk of the narrow ML-estimator of μ is then given by:*

$$\mathrm{br}_{\mathrm{narr}} = Eb'Kb + E\tau_0^2.$$

Proof: Compute the expectation of the risk function from Corollary 2.4.2 by first conditioning on θ. □

Now recall the conclusions reached on p. 73: It was shown that from the viewpoint of the risk function, the narrow and wide ML-estimators would approximate the Bayes estimator, as the variance of a normal prior tended to infinity or zero. It would be reasonable to expect an analogous conclusion to be valid for the Bayes risk. This is indeed true and is easily seen from the above corollaries. We shall give the argument for the case of prior variance tending to zero: Let $K = kK_0$ as on p. 73, and consider the expression for the Bayes risk of the Bayes estimator given by Corollary 4.3.2:

$$\begin{aligned}
\mathrm{br}_{\mathrm{Bayes}} &= Eb'\left[(J^{22})^{-1} + \frac{1}{k}K_0^{-1}\right]^{-1} b + E\tau_0^2 \\
&= Ekb'\left[k(J^{22})^{-1} + K_0^{-1}\right]^{-1} b + E\tau_0^2 \\
&\underset{k\to 0}{\longrightarrow} E\tau_0^2.
\end{aligned}$$

We are assuming sufficient regularity to use Lebesgue's dominated convergence theorem. Now consider the Bayes risk of the narrow ML-estimator as given by Corollary 4.3.5:

$$\mathrm{br}_{\mathrm{narr}} = Ekb'K_0b + E\tau_0^2 \underset{k\to 0}{\longrightarrow} E\tau_0^2.$$

We have thus shown that

$$\lim_{k\to\infty} \mathrm{br}_{\mathrm{Bayes}} - \mathrm{br}_{\mathrm{narr}} = 0.$$

The other case, $k \to \infty$, is straightforward.

We summarize our conclusions about the wide and narrow ML-estimators as seen from

a Bayesian point of view:

- In situations with "vague" prior information about the correctness of the narrow model, corresponding to large variance in the prior distribution of δ, the wide ML-estimator is a good approximation to the Bayes estimator.
- In situations with precise prior information about the correctness of the narrow model, corresponding to a prior distribution for δ concentrated around zero, the narrow ML-estimator is a good approximation to the Bayes estimator.

Remark: In the above statements we are primarily thinking of approximation in the sense of close Bayes risk. Using previous results it should not be too difficult, however, to show that the approximation is also good in the sense of close estimators. For example (4.8) on p. 70 immediately gives that the wide ML-estimator is the limiting case of the Bayes estimator when the prior variance tends to infinity.

As we have seen, the narrow and the wide ML-estimators can be good approximations to the Bayes estimator in many cases, but can we do even better? Suppose that we have determined a prior distribution and are looking for an approximation to the Bayes estimator. Let us search the class of compromise estimators for a candidate. We shall restrict the problem to the case with one-dimensional γ. Consider the compromise estimator determined by letting

$$w(\theta, Z) = \frac{\hat{\delta}_{\text{Bayes}}}{Z},$$

where $\hat{\delta}_{\text{Bayes}}$ is the Bayes estimator of δ in the situation with one observation $Z \sim N(0, J^{22})$ and the conditional distribution of δ given θ as prior. (In practice, substitute an estimator for θ as usual.) Comparing Theorems 3.1.1 and 4.1.1 immediately gives that this compromise estimator has the same risk function as the Bayes estimator! And since the risk functions are equal, it follows that the Bayes risks will be equal. We have thus determined a perfect approximation from an asymptotic point of view! (If possible, we do of course still recommend the use of the original Bayes estimator since the finite sample properties will be different.)

To compute the approximating compromise estimator it is only necessary to compute a Bayes estimator in the one observation normal situation. This may be considerably simpler than computing the Bayes estimator of μ based on the original data.

This kind of "Bayes inspired" compromise estimators are actually the kind of Bayes estimators studied in EMMM. But the link to the genuine Bayes estimators of μ is not clarified there.

We have now compared the different estimators to the Bayes estimator, but there is one more question to ask. How do the narrow and wide ML-estimator compare to each other in the Bayesian context? Does there exist some kind of tolerance radius when the criterion is Bayes risk? We shall again consider the situation with a normal prior for δ. Comparison of Corollary 4.3.3 and 4.3.5 immediately gives that the narrow estimator is better than the

wide if and only if

$$Eb'Kb < Eb'J^{22}b. \tag{4.10}$$

If $J^{22} - K$ is almost surely positive definite, narrow estimation is better than wide for all estimands. (Except those with $b = 0$ a.s, for which narrow and wide estimation are equivalent.) Let us now specialize to the situation where γ is one dimensional and J^{22} does not depend on θ. This is in fact the situation in several of our examples. To stress the fact that the prior covariance matrix is scalar, denote K by ρ^2. Now (4.10) specializes to

$$\rho^2 Eb^2 < J^{22} Eb^2.$$

That is, narrow estimation is better than wide estimation for all estimands with b not almost surely equal to 0 if and only if

$$|\rho| < \sqrt{J^{22}},$$

thus providing an analogue to (2.15) on p. 26 in this special situation.

4.4 An example

We shall now consider one of our previous examples in the Bayesian context.

EXAMPLE 2 CONTINUED (TWO EXPONENTIAL VARIABLES)

Assume as before that we are estimating $\mu = 1/(\theta\gamma)$, and recall that $J^{22} = 2$, $b = 1/(2\theta)$ and $\tau_0^2 = 1/(2\theta^2)$. Assume now that we describe our prior knowledge about δ by a $N(0, \rho^2)$ distribution, thus assuming δ to be independent of θ. From Corollary 4.1.1 we then obtain the risk function for the Bayes estimator:

$$\begin{aligned} r_{\text{Bayes}}(\theta, \delta) &= b' \left[(J^{22})^{-1} + \rho^{-2}\right]^{-1} \left[(J^{22})^{-1} + \rho^{-2}\delta\delta'\rho^{-2}\right] \left[(J^{22})^{-1} + \rho^{-2}\right]^{-1} b + \tau_0^2 \\ &= \frac{1 + 2\delta^2/\rho^4}{2\theta^2(1 + 2/\rho^2)^2} + \frac{1}{2\theta^2}. \end{aligned}$$

A comparison between the risk function of some different Bayes estimators and the narrow and wide ML estimator is given in figure 4.1. We have set $\theta = 0.1$, thus plotting the risk functions only as a function of δ. Note how the risk of the Bayes estimators approximates the risk of the wide or narrow ML-estimator when ρ is respectively large or small.

Let us now turn to the Bayes risk of our estimators. The Bayes risk of the Bayes

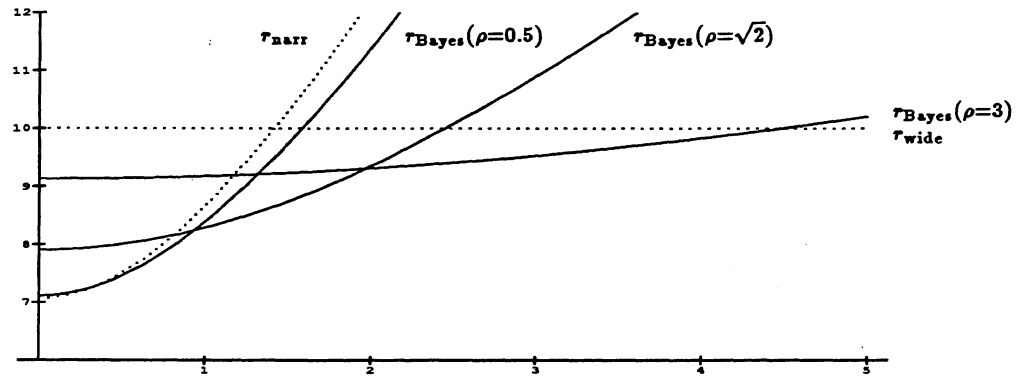

Figure 4.1: Square root of risk functions: Two exponential variables. Narrow and wide risk function shown by dotted lines, and risk functions for Bayes estimators with solid lines. The three Bayes estimators correspond to $\rho = 0.5$, $\rho = \sqrt{2}$ (equal to the tolerance radius) and $\rho = 3$. The estimand is $\mu = 1/(\theta\gamma)$ and $\theta = 0.1$.

estimator is given by Corollary 4.3.2:

$$\begin{aligned} \mathrm{br}_{\mathrm{Bayes}} &= Eb' \left[(J^{22})^{-1} + \rho^{-2} \right]^{-1} b + E\tau_0^2 \\ &= \frac{1+\rho^2}{2+\rho^2} E \frac{1}{\theta^2}. \end{aligned}$$

The last expectation should be computed in the prior distribution of θ. If for example θ is given a gamma distribution with parameters $\alpha > 2$ and β, we obtain

$$E\frac{1}{\theta^2} = \frac{\beta^2}{(\alpha-1)(\alpha-2)}.$$

Note that we did not actually have to specify the entire distribution of θ in order to compute the Bayes risk. It will of course suffice to specify the value of $E1/\theta^2$. In our example $1/\theta^2 = \mathrm{Var}\,(V_i|\theta)$, thus having an intuitive accessible interpretation. It may actually be easier to specify the distribution, or only the expectation, of $1/\theta^2$ directly, than to specify the distribution of θ. [2]

For comparison we also compute the Bayes risk of the narrow and wide ML-estimators. The Bayes risk of the wide estimator is given by Theorem 4.3.3:

$$\mathrm{br}_{\mathrm{wide}} = Eb'J^{22}b + E\tau_0^2 = E\frac{1}{\theta^2}.$$

[2]This is an instance of a more general point. In order to meaningfully specify a prior distribution, it will usually be necessary to find a parameterization where the parameter has some kind of interpretation that is accessible to the human mind.

And from Corollary 4.3.4 we obtain the Bayes risk of the narrow estimator:

$$\mathrm{br}_{\mathrm{narr}} = Eb'\rho^2 b + E\tau_0^2 = \left(\frac{\rho^2}{4} + \frac{1}{2}\right) E\frac{1}{\theta^2}.$$

In figure 4.2 we have made a numerical comparison between the Bayes risks of the Bayes estimator and the two ML-estimators. We have plotted the Bayes risks as functions of ρ. The prior expectation of $1/\theta^2$ is chosen to be 100.

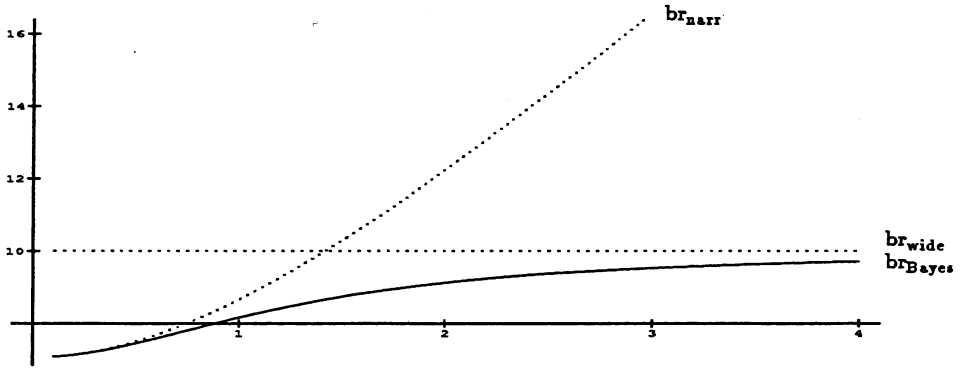

Figure 4.2: Square root of Bayes risks plotted as a function of ρ : Two exponential variables. Bayes risk for narrow and wide ML-estimator shown with dotted lines, and Bayes risk for the Bayes estimator shown with a solid line. The estimand is $\mu = 1/(\theta\gamma)$ and $E1/\theta^2 = 100$.

Note that the narrow ML-estimator is close to the Bayes estimator for small ρ and the wide ML-estimator is close to the Bayes estimator for large ρ. (Cf. the comments on p. 76.) Indeed it seems that for many purposes the narrow ML-estimator would be an acceptable approximation for $\rho < \sqrt{2}$, and the wide one would be acceptable for $\rho > \sqrt{2}$. (In this situation the tolerance radius is $\sqrt{2}$, cf. the remark on p. 77.) We have thus proposed calculationally simple approximations to the Bayes estimator. (Compare the expressions for the ML-estimators, as given by (2.19) and (2.18) on p. 34, to the fact that the Bayes estimator would have to be determined by numerical integration in this situation.)

One final remark is in order: In Chapter 2 we obtained a criterion based on the risk function that told us to use the narrow ML-estimator instead of the wide one if and only if $\delta < \sqrt{2}$. Here we obtained the analogous criterion: Use the narrow ML-estimator instead of the wide one if and only if $\rho < \sqrt{2}$. It may seem as if we have gone to a great deal of extra theory and achieved very little new. This is not true. There is a major difference between the two criteria. The point is that the first one can not be assessed since it depends on the unknown quantity δ, while the second can easily be employed in practice once the prior beliefs have been stated. □

In a similar way we could analyze the other examples considered in the Bayesian context. If we stayed with the normal priors for δ, the derivation of the risk function of the Bayes estimator would be simple in all examples. Other priors would normally require numerical

integration. To compute the Bayes risk from the risk function, we would have to resort to numerical integration in most situations.

4.5 Conclusions

I do of course not believe that there is one easy and complete answer to the questions raised in the introduction. I shall anyway take the chance and sum up what I feel are the conclusions of the investigations made in this thesis:

- If it is not computationally too difficult, one should use the wide model, quantify any prior knowledge as a prior distribution, and use the resulting Bayes estimator.
- If there are computational difficulties, one might consider to use either the wide or the narrow ML-estimator as an approximation to the Bayes estimator. The loss incurred, as measured by Bayes risk, could be approximated by methods of this chapter. An even better solution would be to determine a compromise estimator giving a closer approximation to the Bayes estimator than both ML-estimators. (Cf. the discussion on p. 76.)

References

Berger, J.O. (1982). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. In *Statistical Decision Theory and Related Topics III,* eds. Berger and Gupta, 109-141. Academic press, New York.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Second Edition. Springer-Verlag, New York.

Bickel, P.J. (1984). Parametric robustness: Small biases can be worthwhile. *Annals of Statistics* **12**, 864-879.

Billingsley, P. (1986). *Probability and Measure.* Second edition, Wiley, New York.

Hjort, N.L. (1980). *Kompendium S 205: Sannsynlighetsregning III,* University of Oslo.

Hjort, N.L. (1986). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scandinavian Journal of Statistics* **13**, 63-85.

Hjort, N.L. (1991a). Estimation in moderately misspecified models. Technical report, University of Oslo; submitted for publication.

Hjort, N.L. (1991b). The exact amount of t-ness that the normal model can tolerate. *Journal of the American Statistical Association,* 1992, to appear.

Lehmann, E.L. (1983). *Theory of Point Estimation.* Wiley, Singapore.

Neuhaus, W. (1985). Choice of statistics in linear Bayes estimation, *Scandinavian Actuarial Journal, 1-26.*